We will start at 2:05 pm!

Thanks for coming early!

Yesterday

Fundamental

1. Value of visualization
2. Design principles
3. Graphical perception

Record Information

Copyright, 1878, by MUYBRIDGE.
MORSE'S Gallery, 417 Montgomery St., San Francisco.
THE HORSE IN MOTION.

Support Analytical Reasoning

Communicate Information to Others

DIAGRAM OF THE CAUSES OF MORTALITY IN THE ARMY IN THE EAST.

2. APRIL 1855 TO MARCH 1856.

1. APRIL 1854 TO MARCH 1855.

JUNE
MAY
JULY
AUGUST
APRIL 1855
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER
JANUARY 1856
FEBRUARY
MARCH
JULY
AUGUST
JUNE
MAY
BULGARIA
SEPTEMBER
APRIL 1854
CRIMEA
OCTOBER
NOVEMBER
DECEMBER
JANUARY 1855
FEBRUARY
MARCH 1855.

The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Novr. 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855; the black area coincides with the red;

Yesterday

Fundamental

1. Value of visualization
2. Design principles
3. Graphical perception

Graphical Integrity

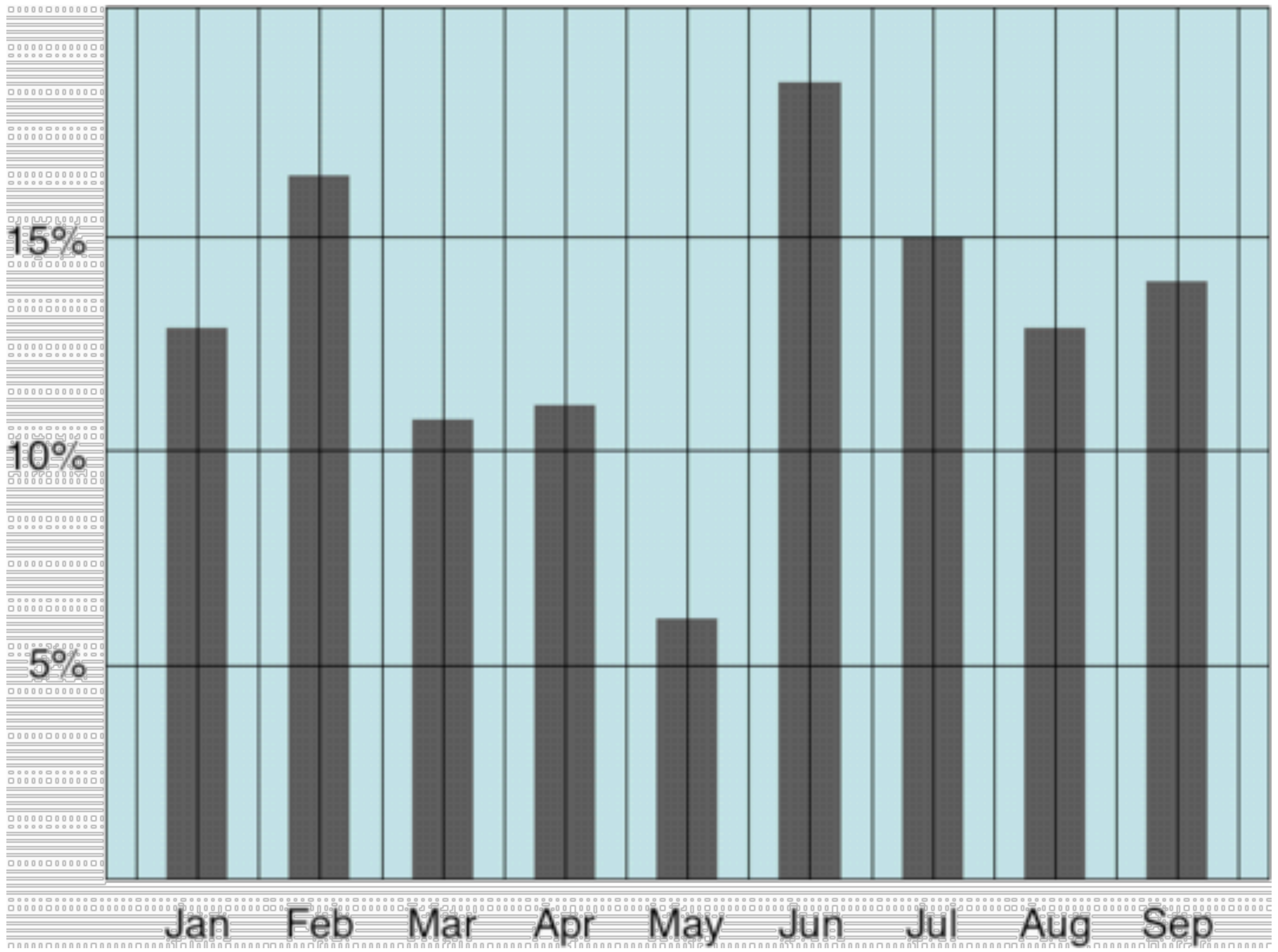
IF BUSH TAX CUTS EXPIRE
TOP TAX RATE
42%
40
38
36
34
39.6%
35%
NOW
JAN. 1, 2013

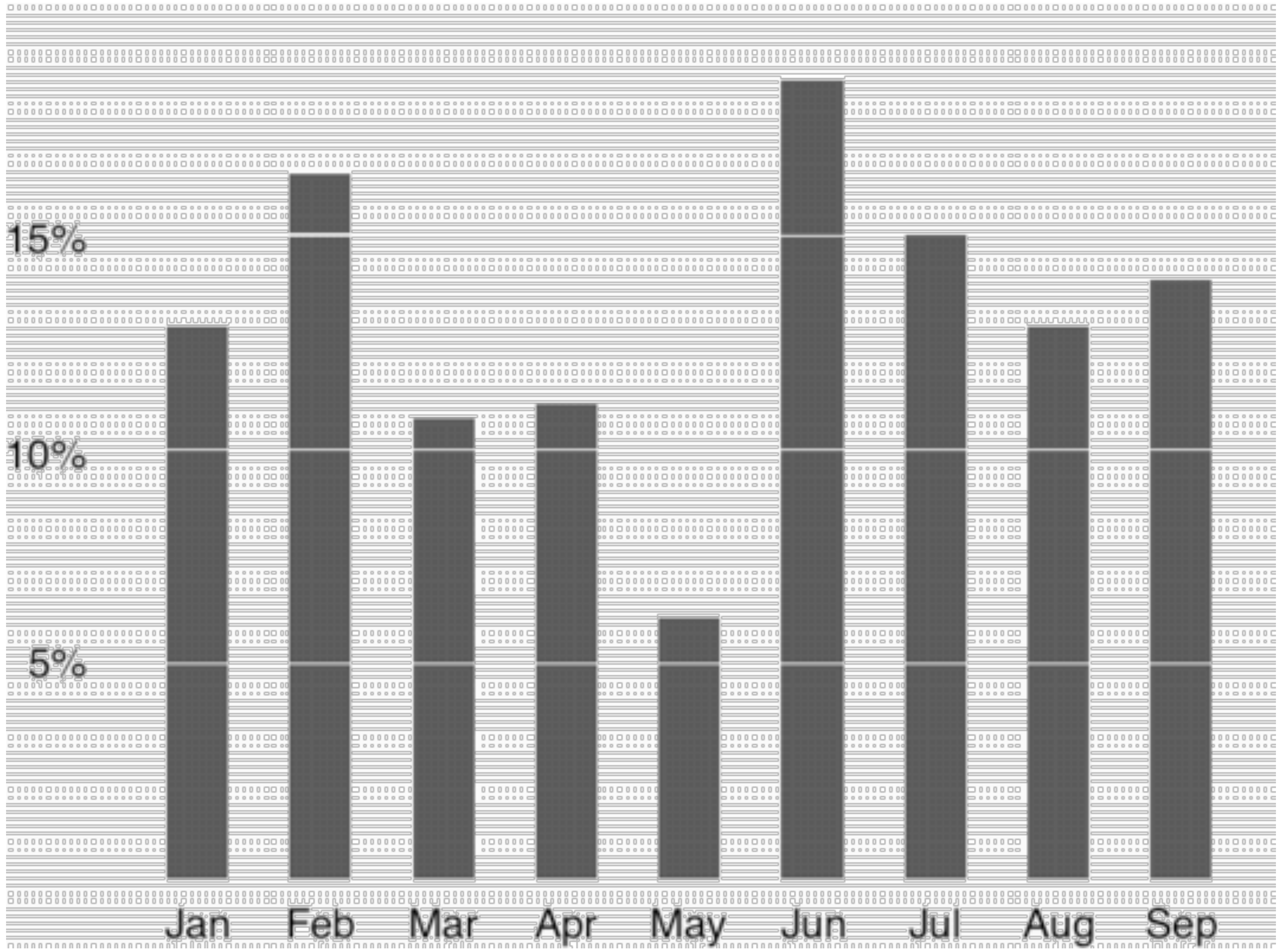
Be mindful about the baseline!

ne, representing 18 miles per
in 1978, is 0.6 inches long.
Fuel Economy Standards for Autos
Set by Congress and supplemented by the Transportation Department. In miles per gallon.
1978
'79
'80
'81
'82
'83
'84
'85
18
19
20
22
24
26
27
27½
ne, representing 27.5 miles per
in 1985, is 5.3 inches long.

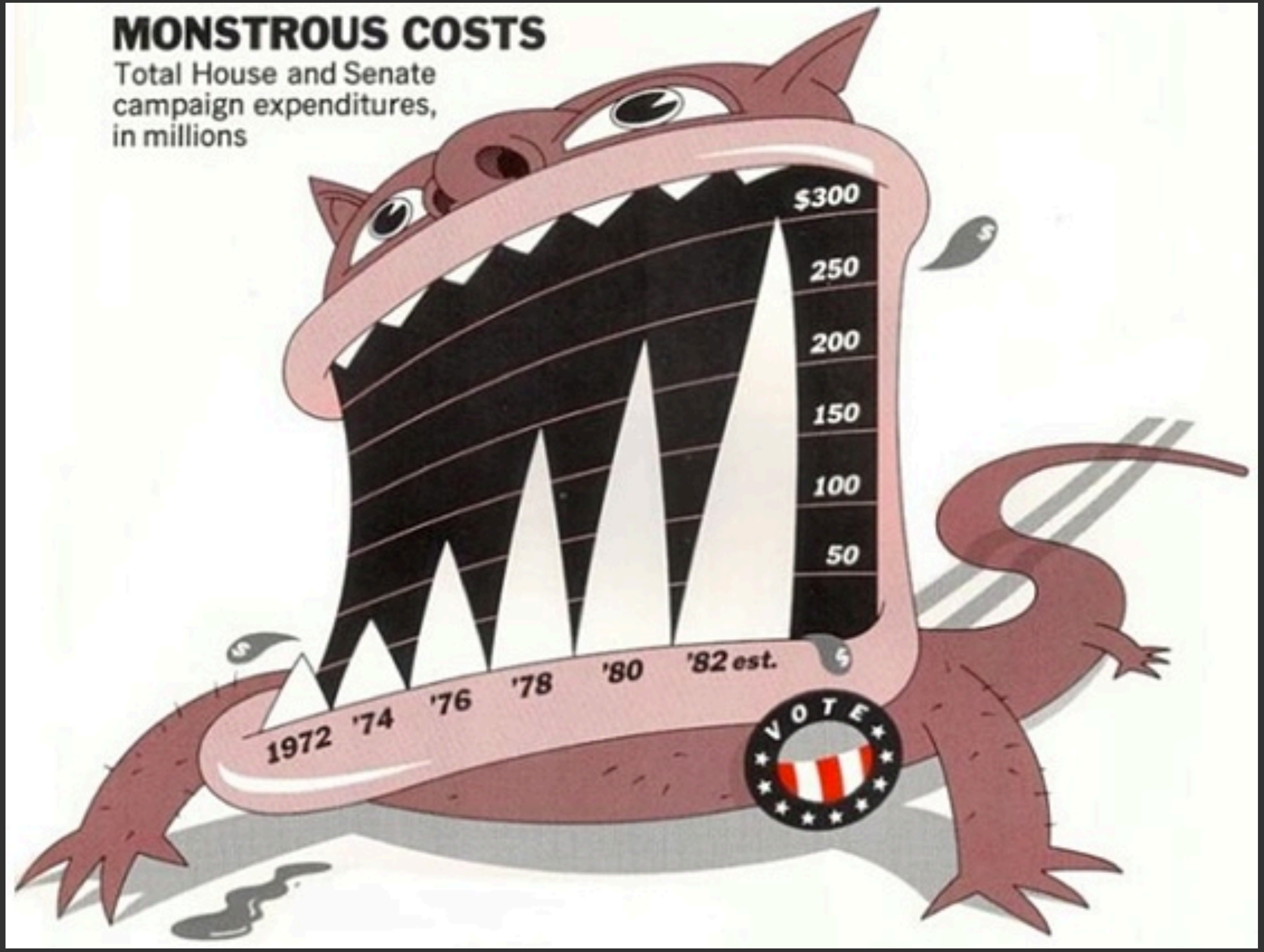
$$\text{Lie Factor} = \frac{\text{Size of effect in graphic}}{\text{Size of effect in data}}$$

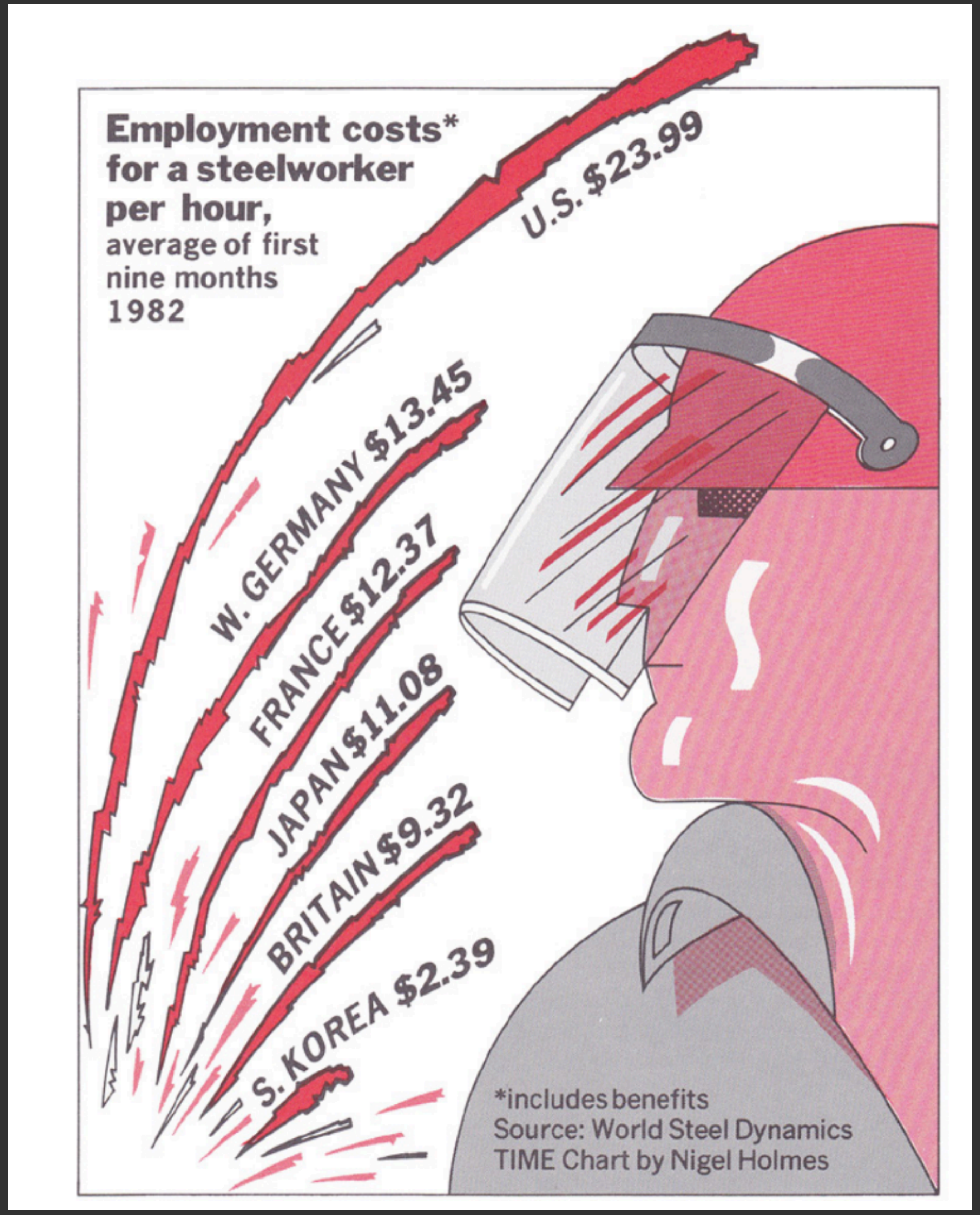
Maximize Data-Ink Ratio



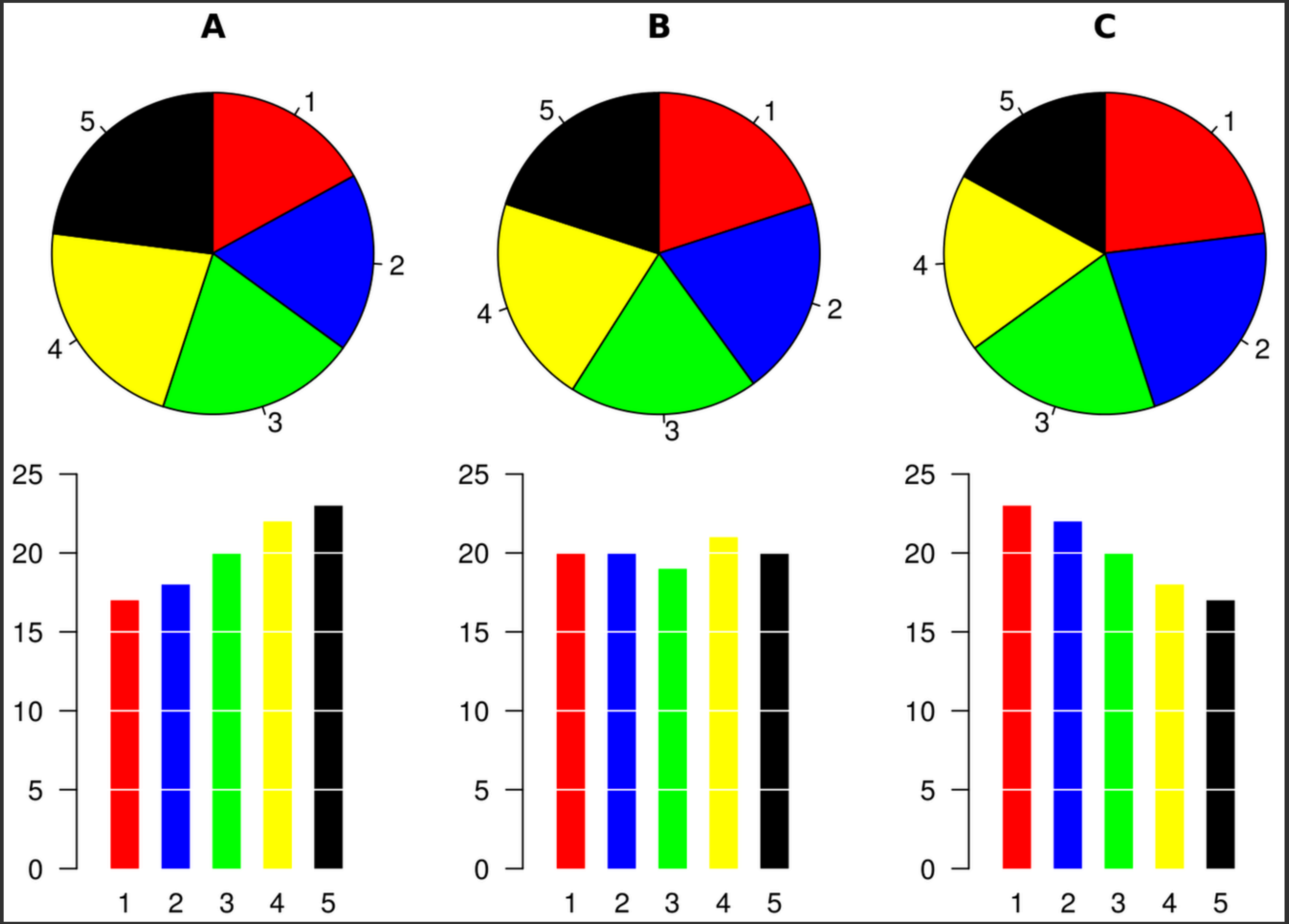


Useful Chart Junks?

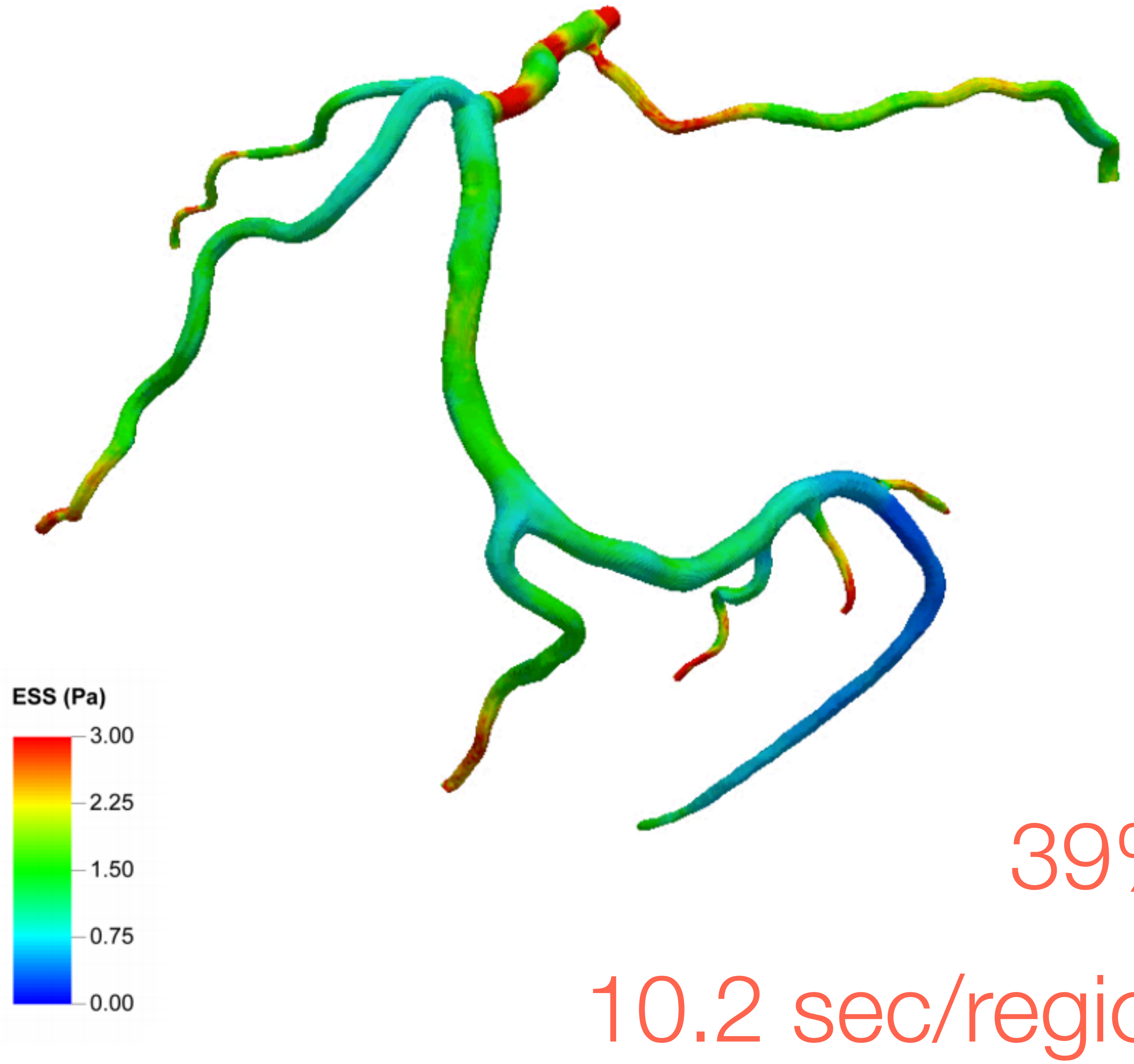


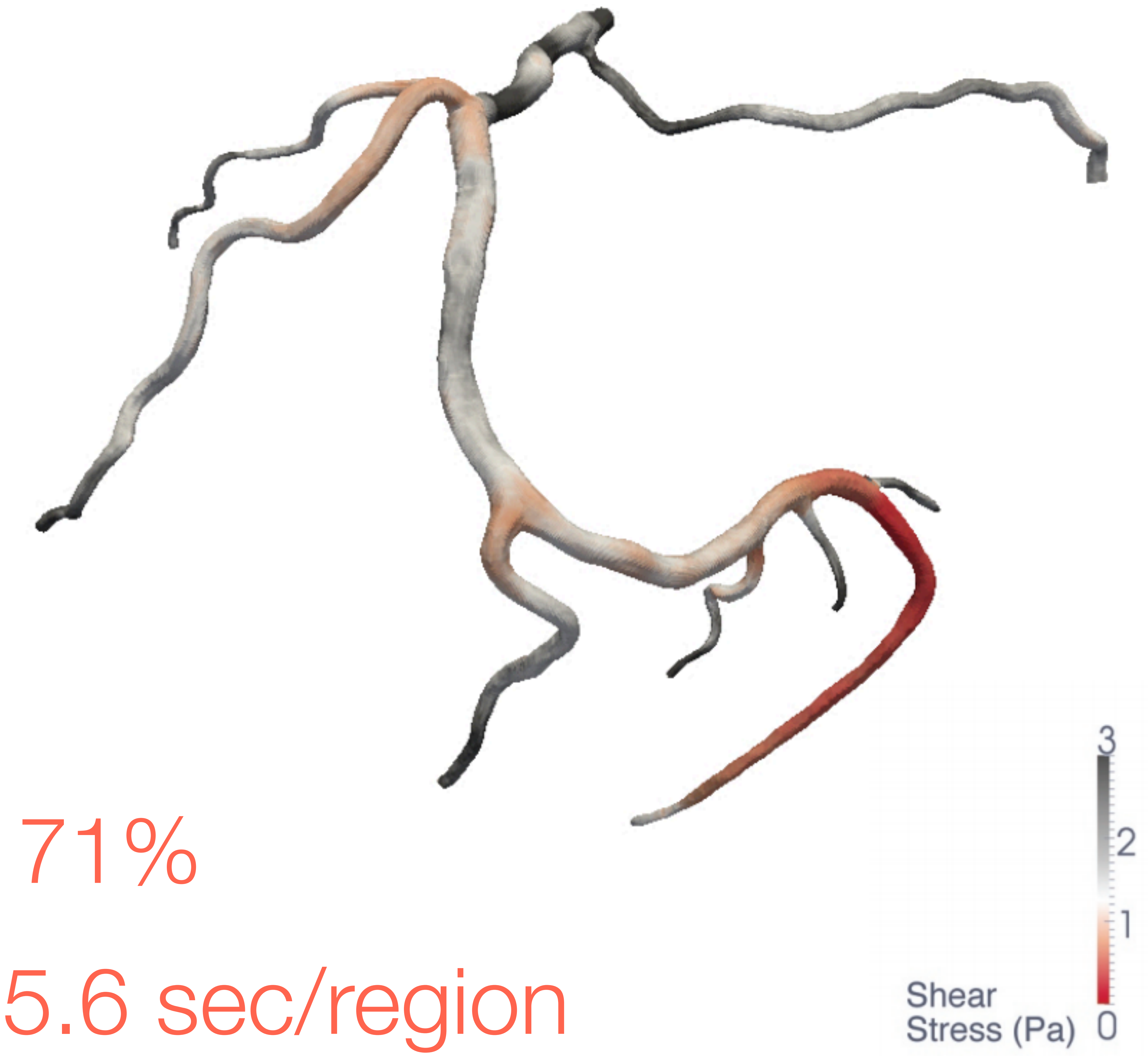


Issues with Pie Charts



Problem with Rainbow Colormap



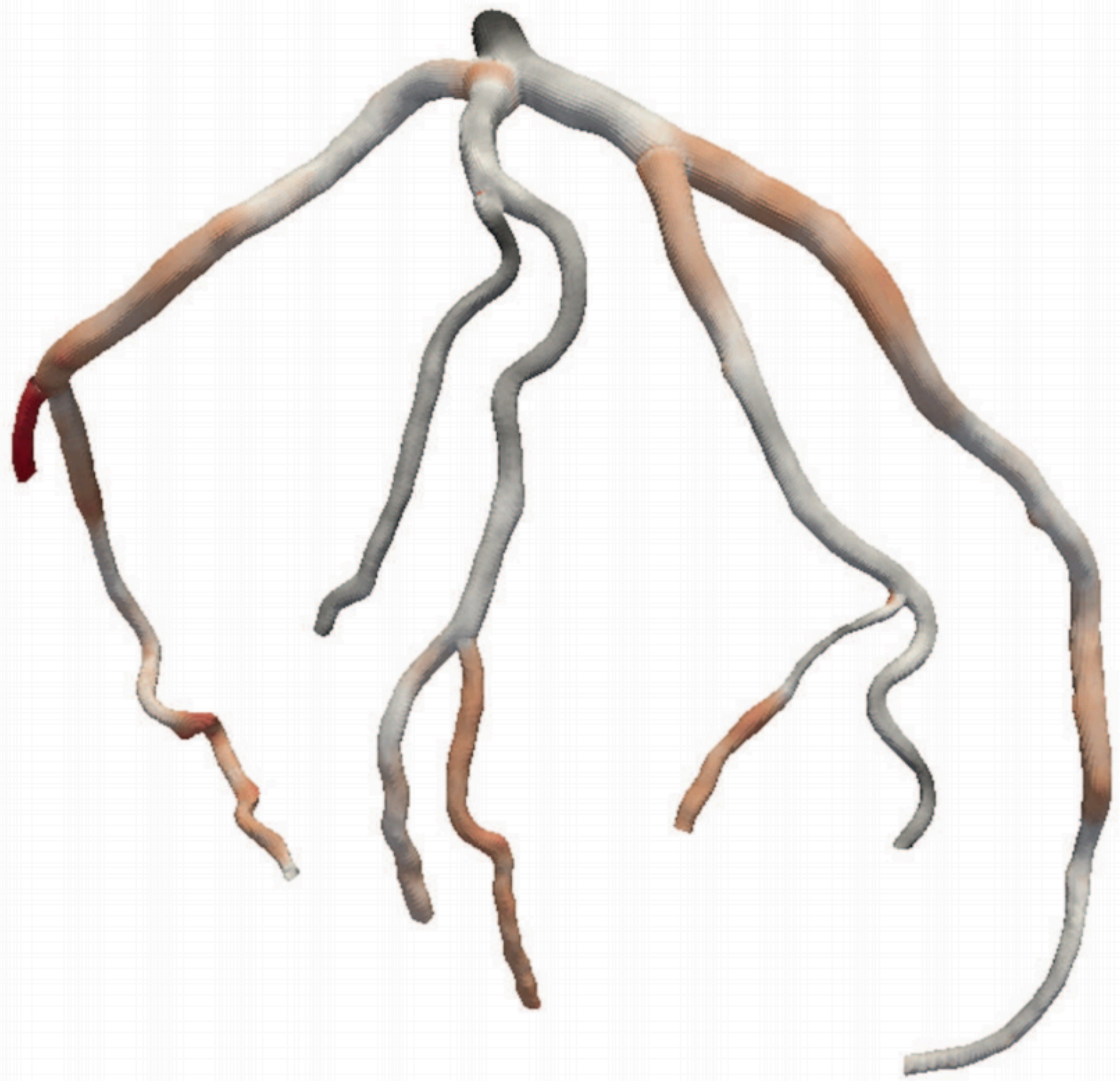


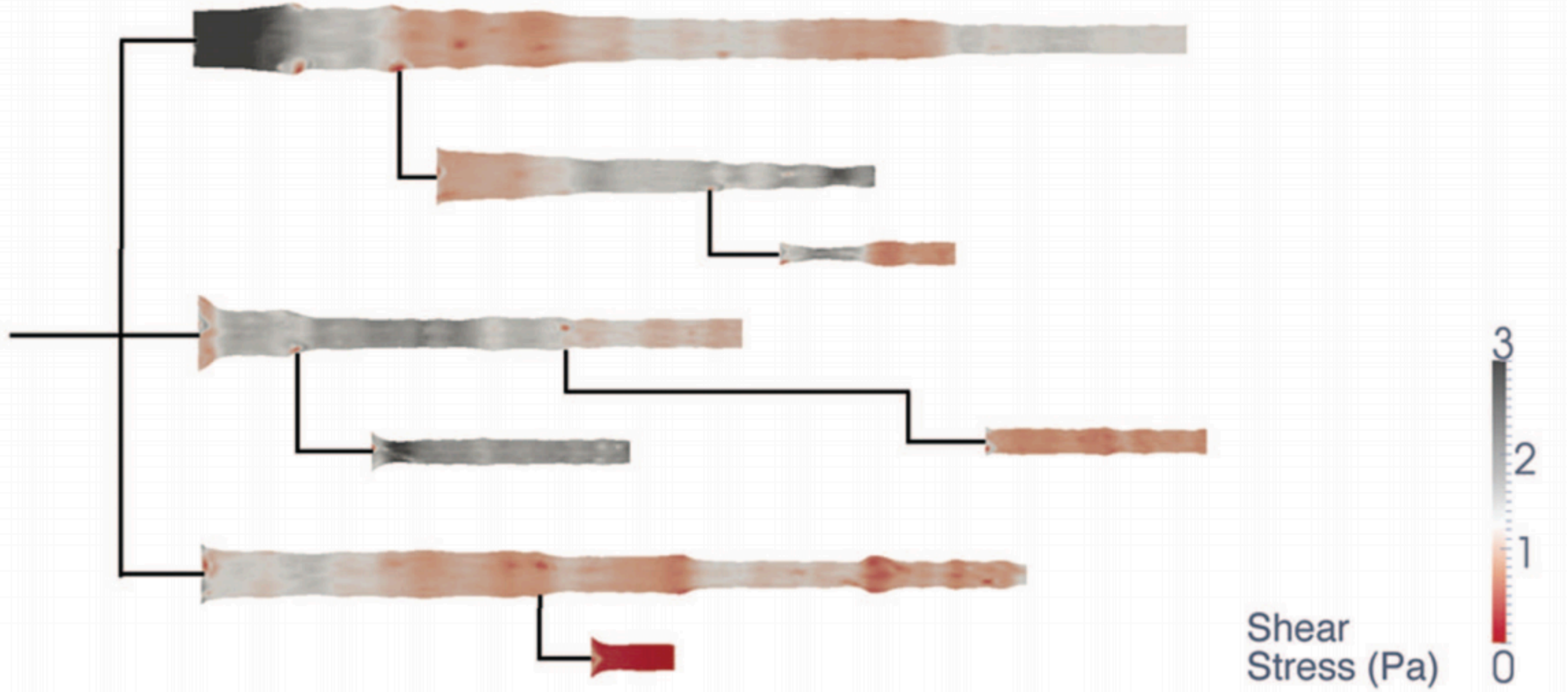
39% → 71%

10.2 sec/region → 5.6 sec/region

[M. Borkin et al 2011]

Problems with 3D Charts





71% ⟶ **91%**

5.6 sec/region ⟶ **2.4 sec/region**

[M. Borkin et al 2011]

Yesterday

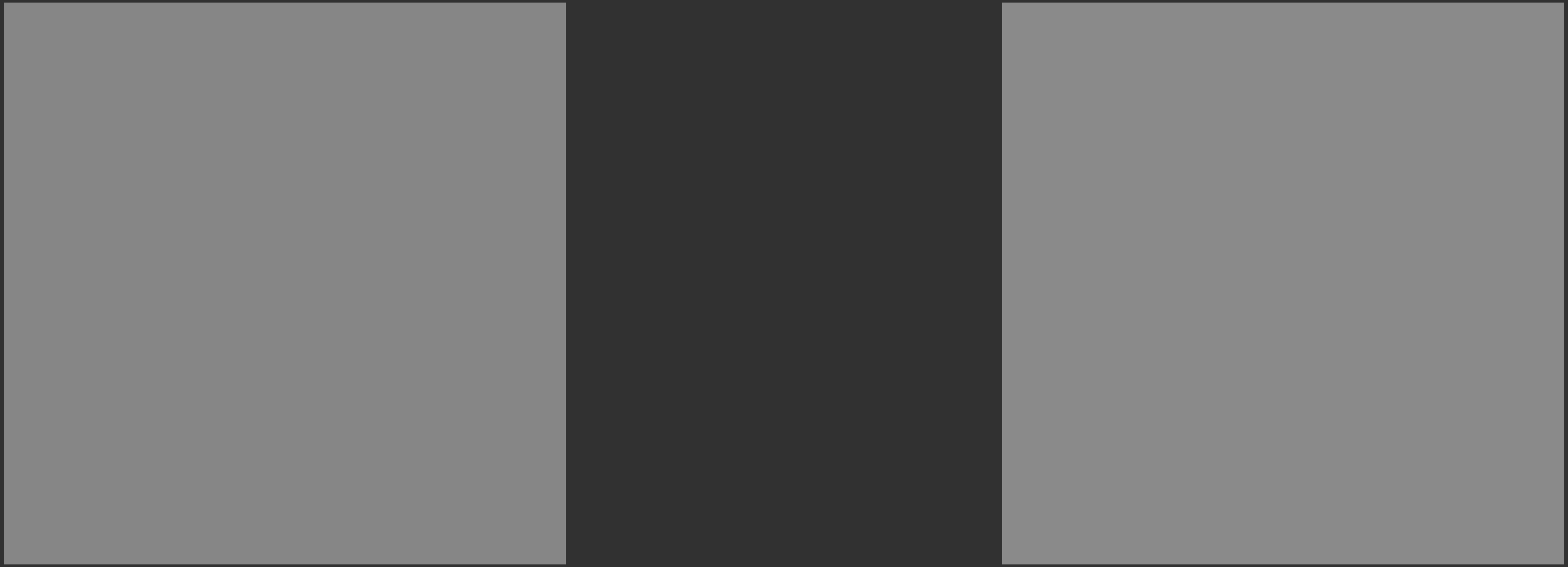
Fundamental

1. Value of visualization

2. Design principles

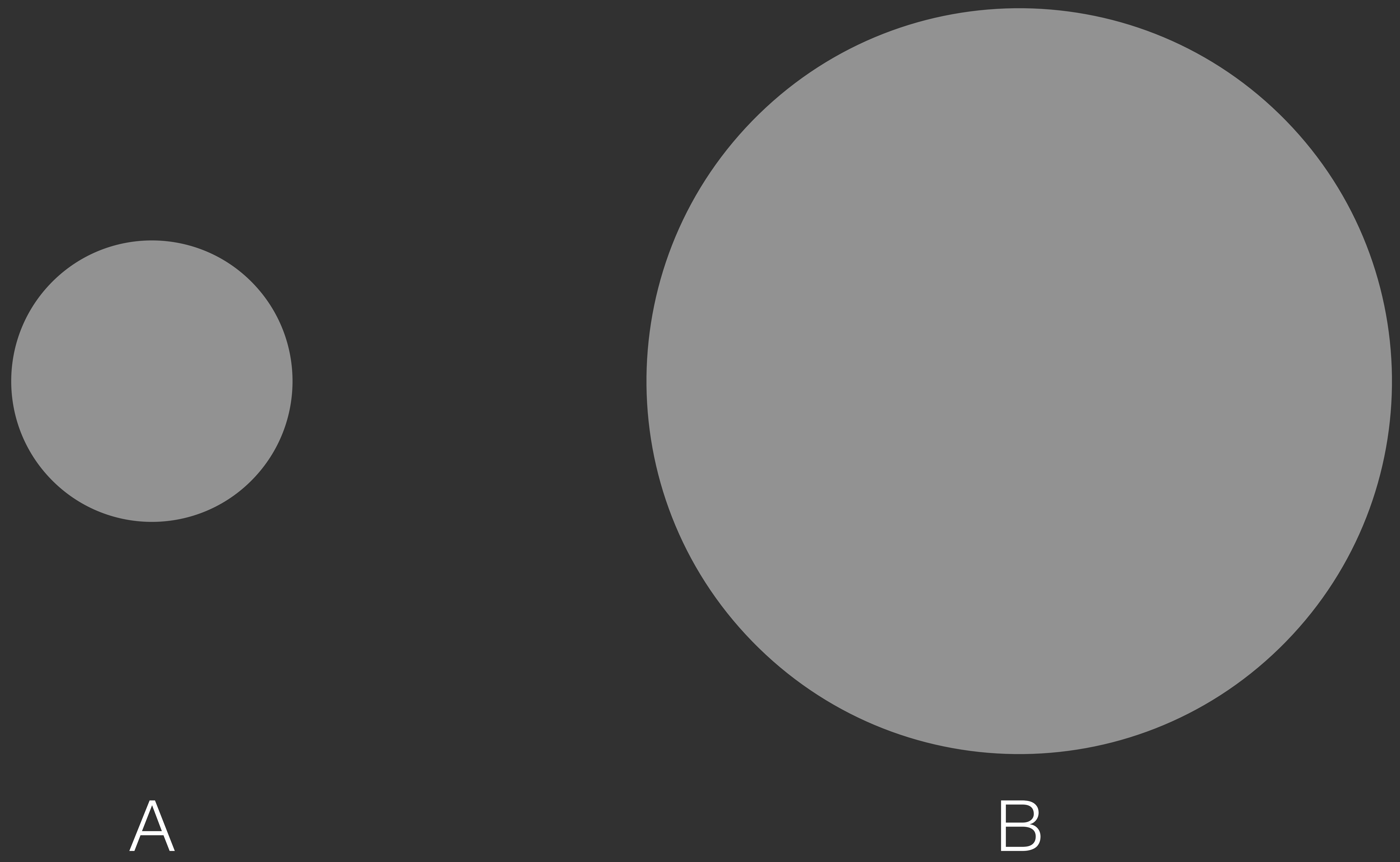
3. Graphical perception

Signal Detection



Which is brighter?

Magnitude Estimation



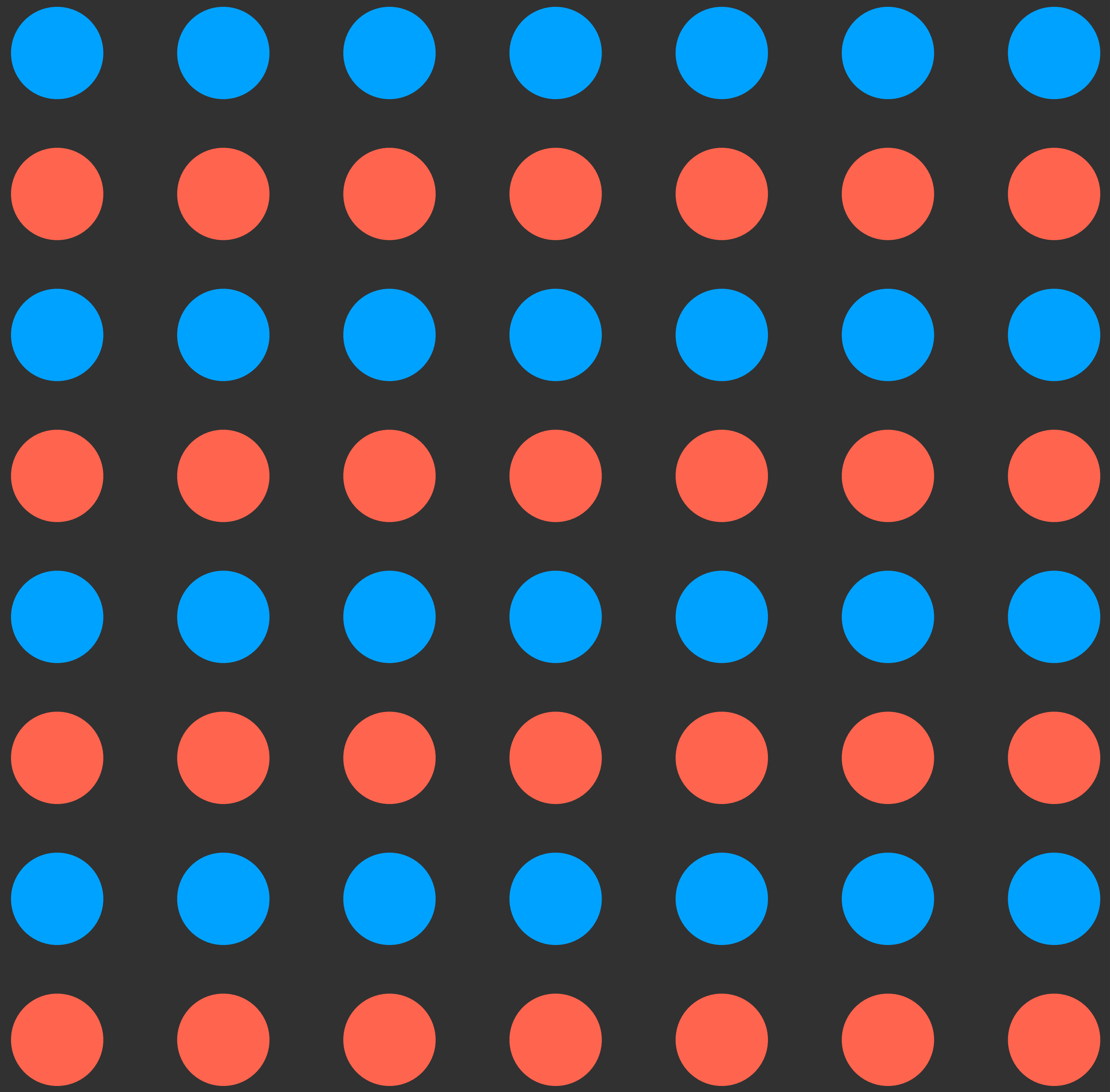
Pre-attentive processing

How Many 3’s?

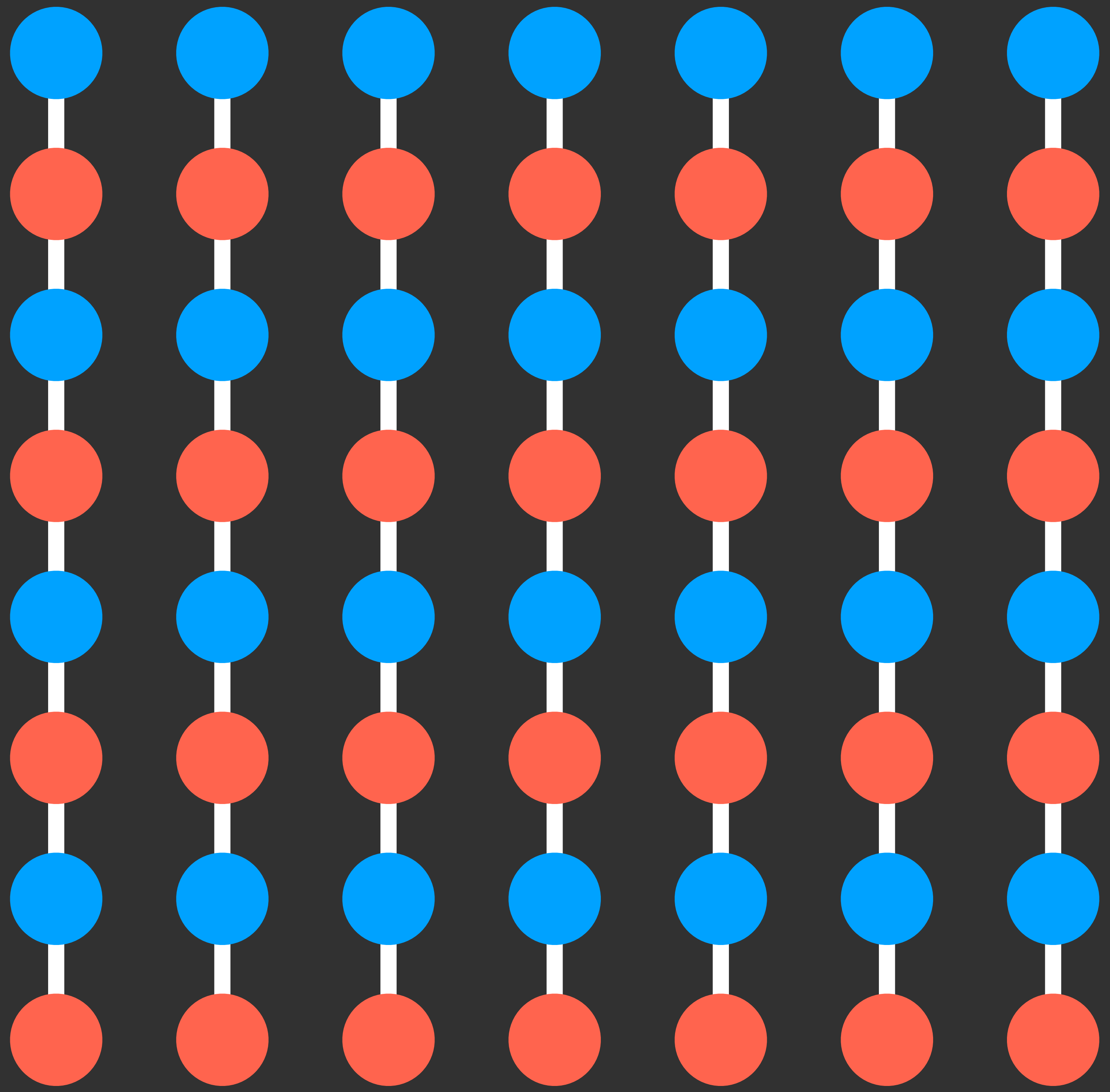
1281768756138976546984506985604982826762
9809858458224509856458945098450980943585
9091030209905959595772564675050678904567
8845789809821677654876364908560912949686

1281768756138976546984506985604982826762
9809858458224509856458945098450980943585
9091030209905959595772564675050678904567
8845789809821677654876364908560912949686

Gestalt Principles



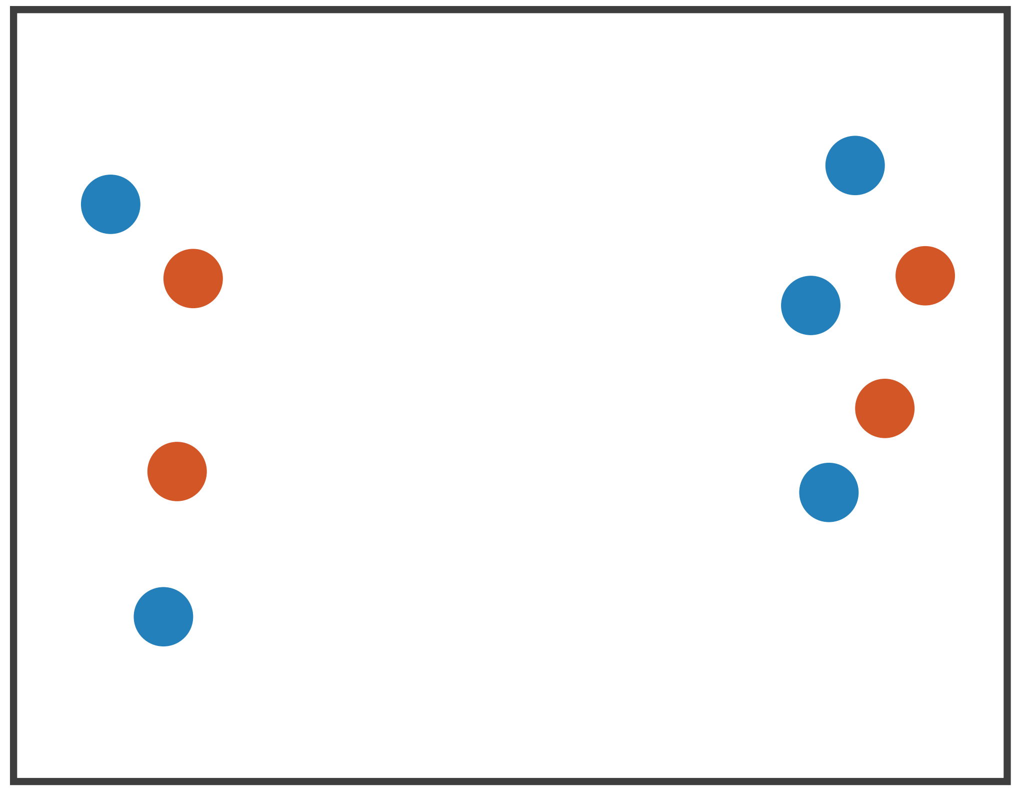
Color **Similarity**



Connection lines

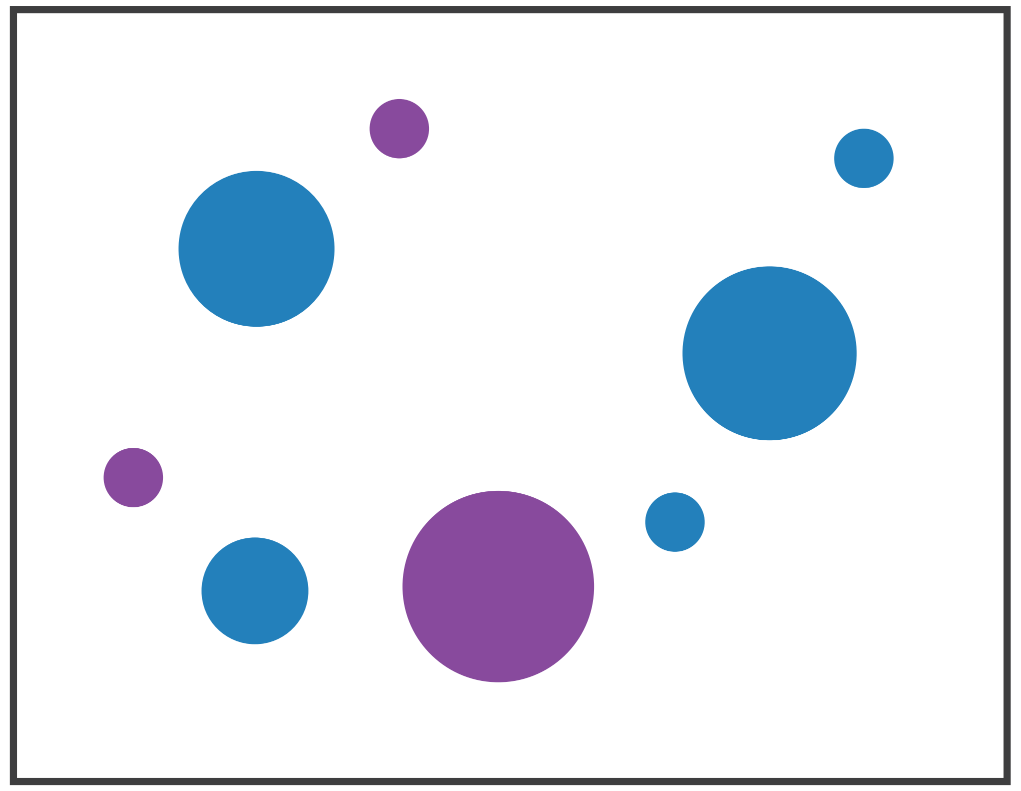
Separability vs. Integrality

Position
+ Hue (Color)



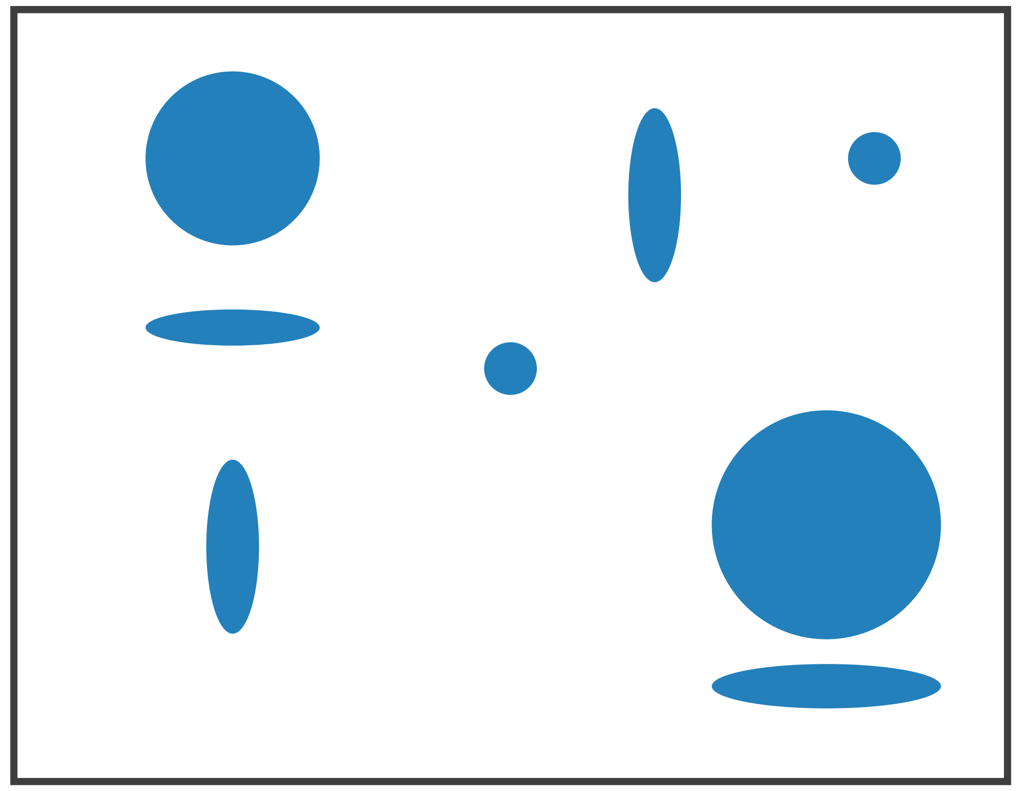
Fully separable

Size
+ Hue (Color)



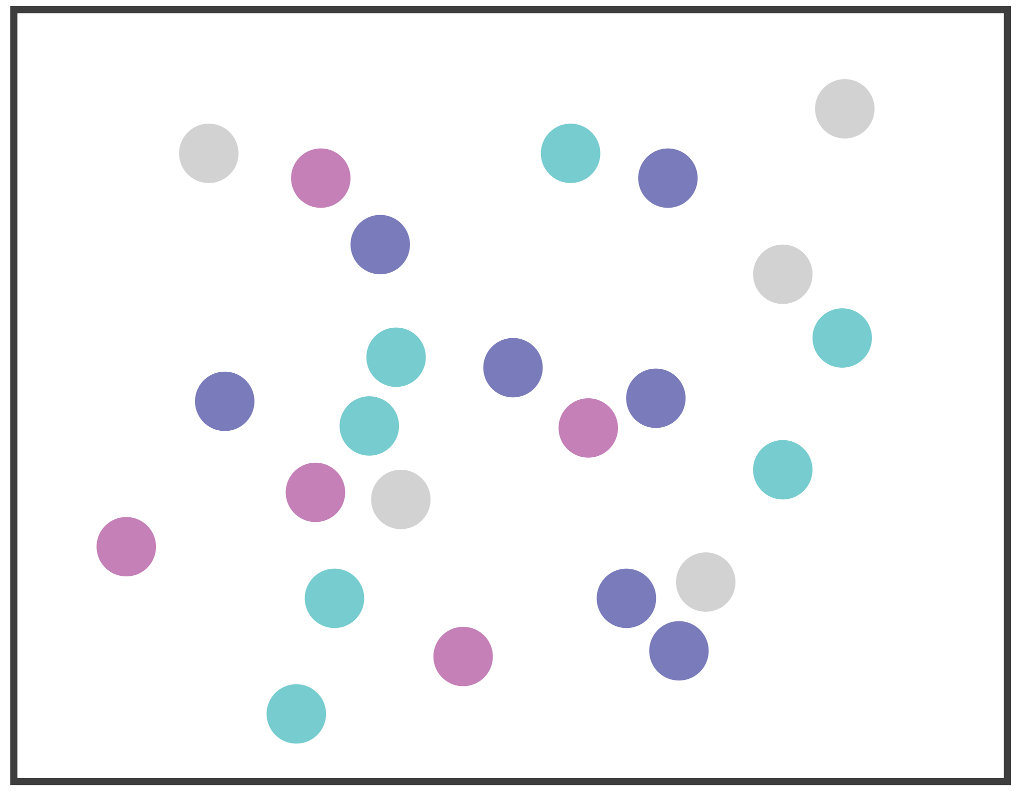
Some interference

Width
+ Height



Some/significant interference

Red
+ Green



Major interference

What we perceive:

2 groups each

2 groups each

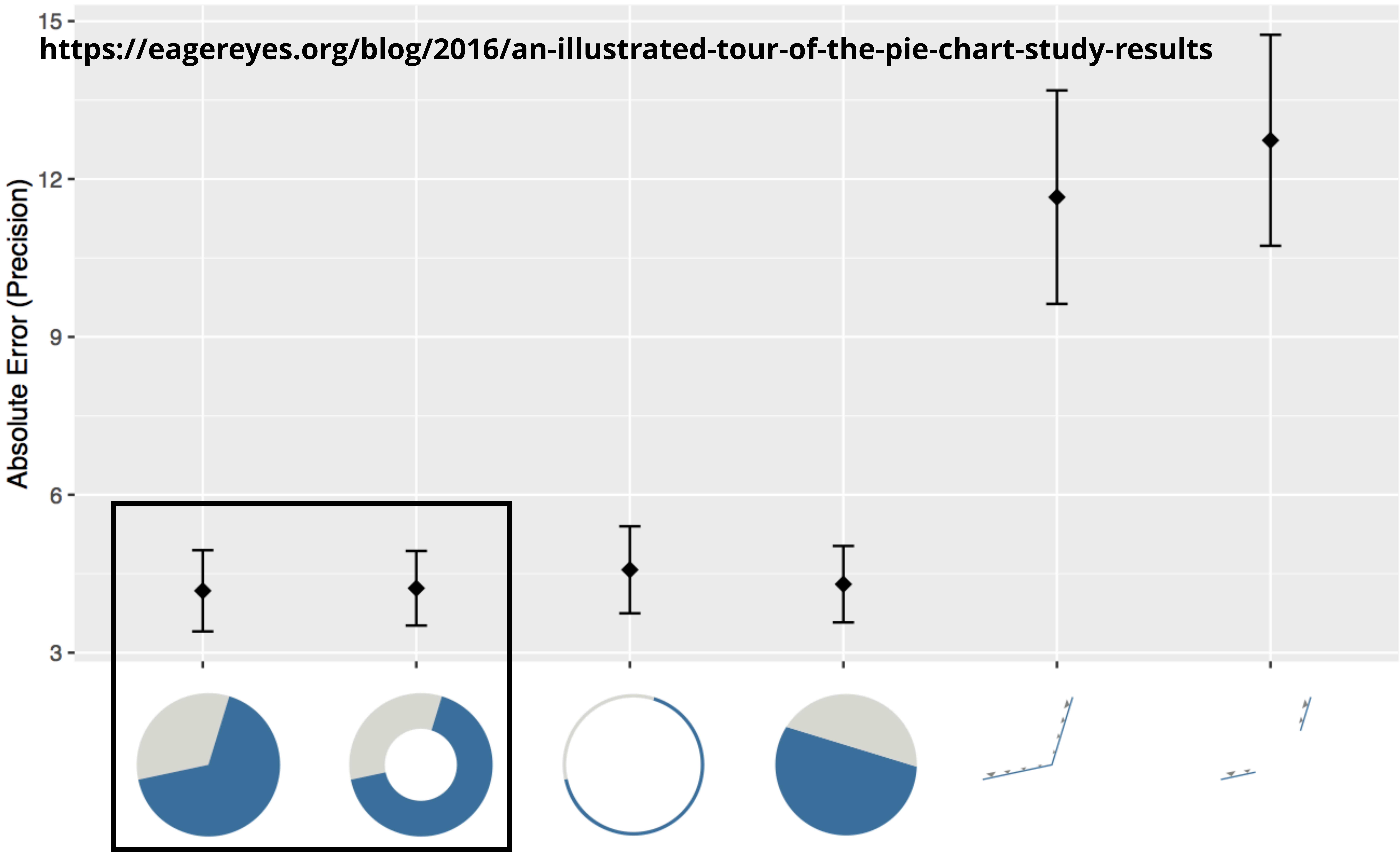
3 groups total: integral area

4 groups total: integral hue

[Tamara Munzner 14]

Change Blindness

http://www.psych.ubc.ca/~rensink/flicker/download/



https://eagereyes.org/blog/2016/an-illustrated-tour-of-the-pie-chart-study-results

Today

Practical

1. Data model and visual encoding
2. Exploratory data analysis
3. Storytelling with data
4. Advanced visualizations

Data Model & Visual Encoding

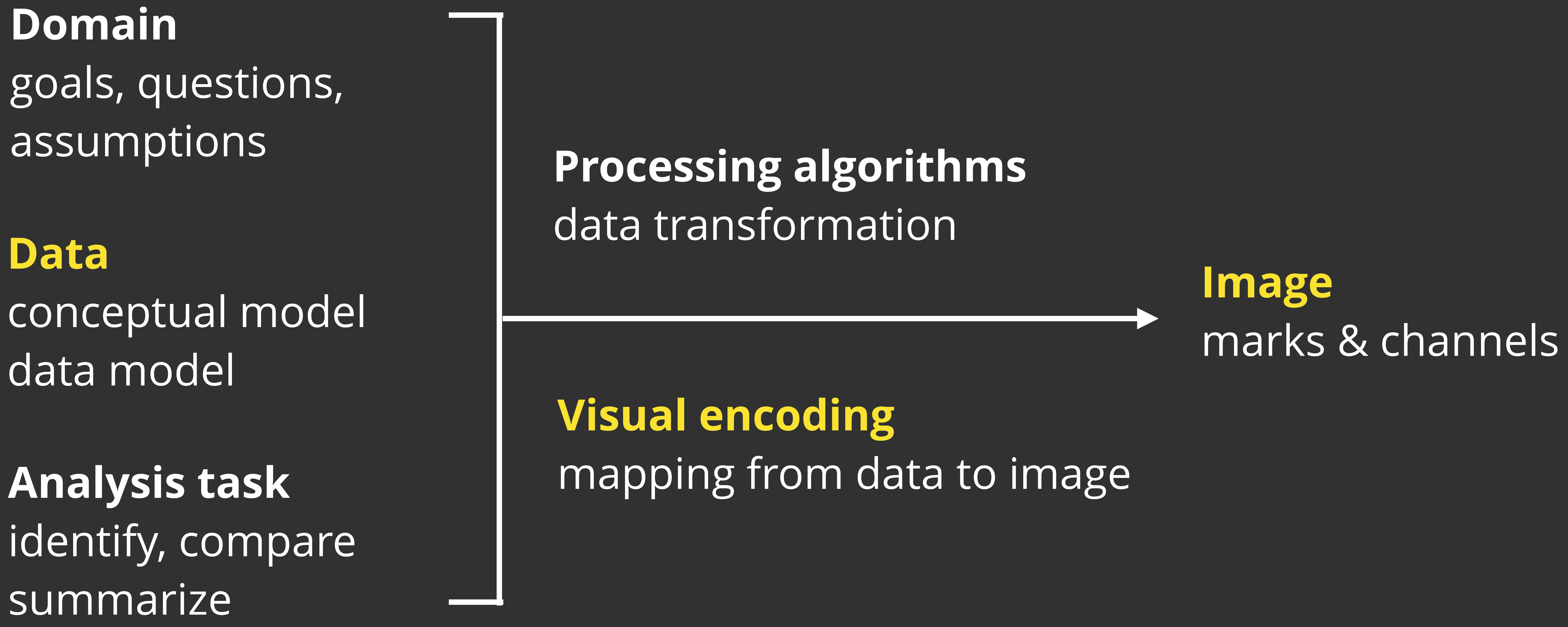
Nam Wook Kim

Mini-Courses — January @ GSAS
2018

Goal

Learn how data
is mapped to images

The Big Picture



Domain
goals, questions,
assumptions

Data
conceptual model
data model

Analysis task
identify, compare
summarize

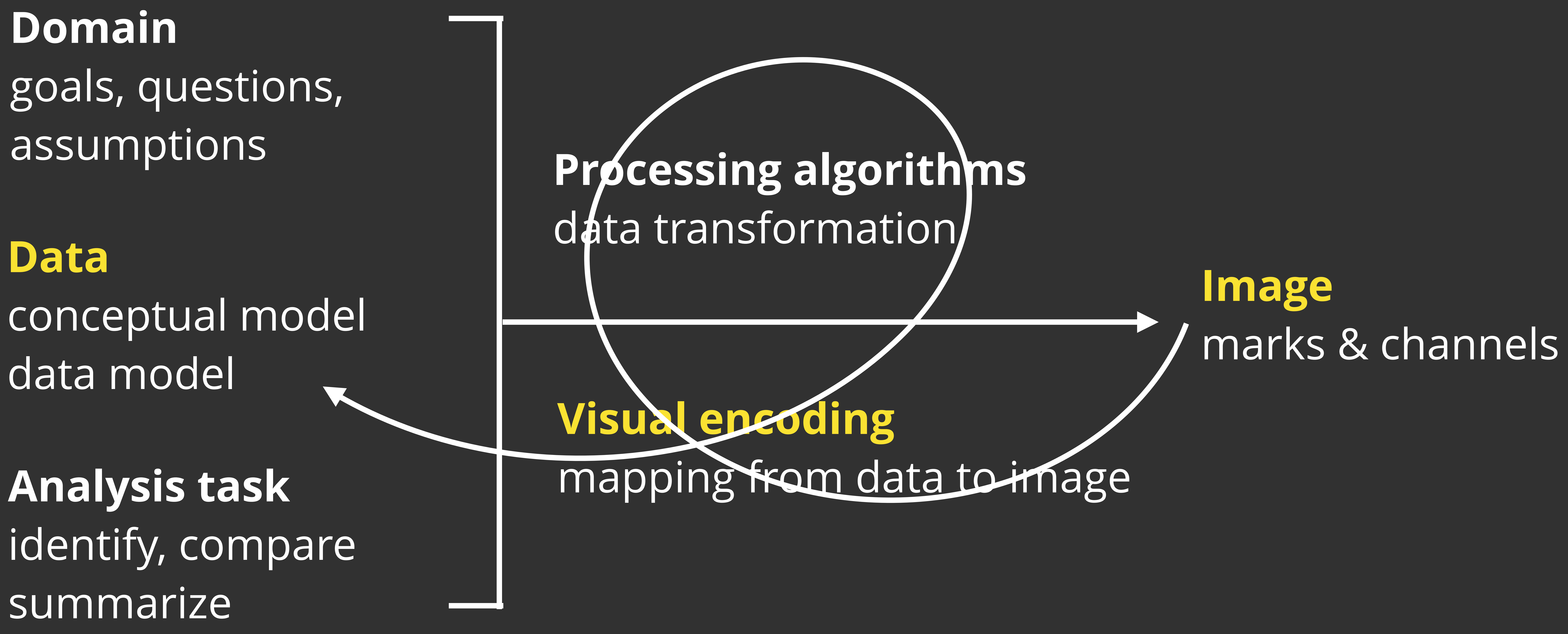
Processing algorithms
data transformation

Visual encoding
mapping from data to image

Image
marks & channels

[Slides from J. Heer]

The Big Picture



Domain
goals, questions,
assumptions

Data
conceptual model
data model

Analysis task
identify, compare
summarize

Processing algorithms
data transformation

Visual encoding
mapping from data to image

Image
marks & channels

[Slides from J. Heer]

Topics

- Data Models
- Image Models
- Visual Encoding
- Formalizing Design

Data Models

Data Models/Conceptual Models

- **Conceptual Models** are mental constructions of the domain

  Include semantics and support reasoning

- **Data Models** are formal descriptions of the data

  Derives from a conceptual model.

  Include dimensions & measures.

- Examples (data vs. conceptual)

  Decimal number vs. temperature

  Longitude, latitude vs. geographic location

Taxonomy of Datasets

1D (sets and sequences)

Temporal

2D (maps)

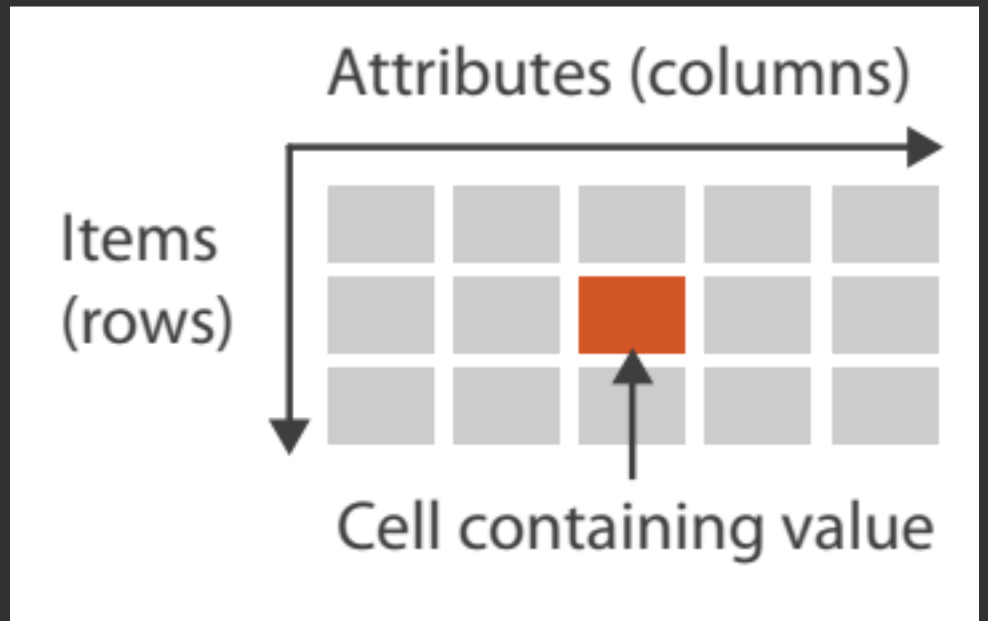
3D (shapes)

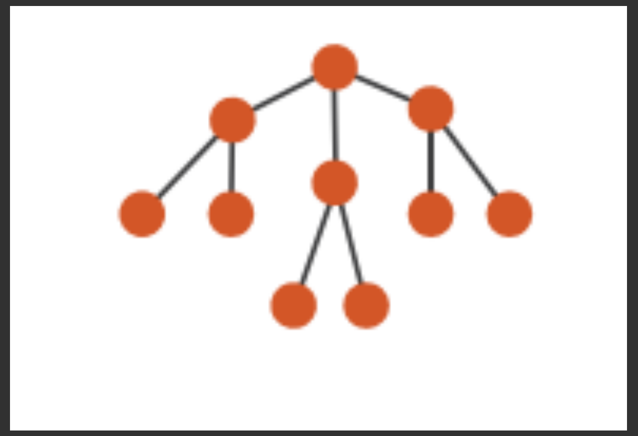
nD (relational)

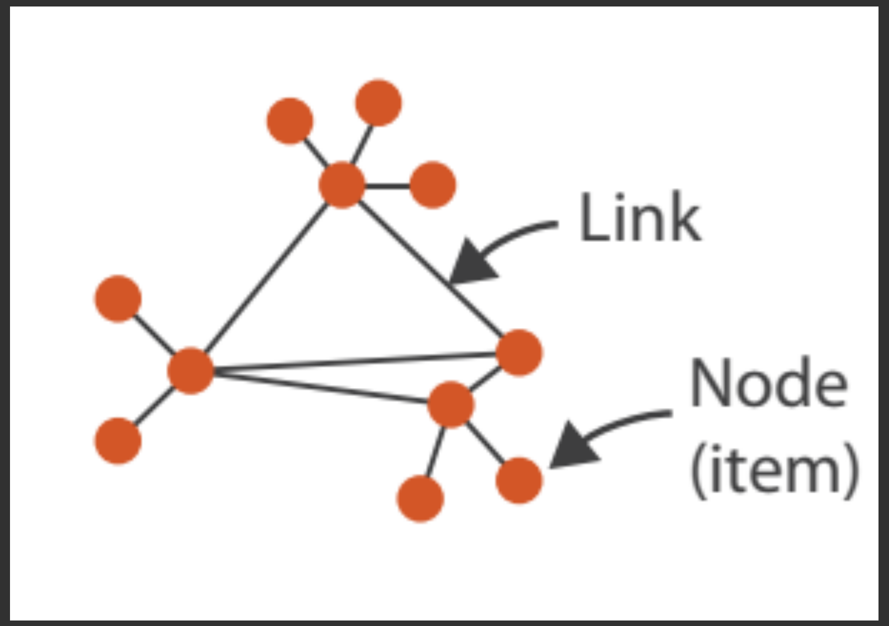
Trees (hierarchies)

Networks (graphs)

and combinations…







[Shneiderman 96]

Data (Measurement) Scales

N—Nominal

O—Ordinal

Q—Quantitative

Data Scales

N—Nominal (labels or categories)

Fruits: apples, oranges, …

Data Scales

N—Nominal (labels or categories)

Fruits: apples, oranges, ...

O—Ordinal

Rankings: 1st, 2nd, 3rd...

Data Scales

N—Nominal (labels or categories)

  Fruits: apples, oranges, ...

O—Ordinal

  Rankings: 1st, 2nd, 3rd...

Q—Quantitative

  Interval (location of zero arbitrary)

    Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)

    Only differences (i.e. intervals) are compared

Data Scales

N—Nominal (labels or categories)

Fruits: apples, oranges, ...

Note: **Q** ⊂ **O** ⊂ **N**

O—Ordinal

Rankings: 1st, 2nd, 3rd...

Q—Quantitative

Interval (location of zero arbitrary)

Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)

Only differences (i.e. intervals) are compared

Ratio (zero fixed)

Physical measurement: length, amounts, counts

Allow direct comparisons like twice as long

Data Scales

Operations

N—Nominal (labels or categories)

Fruits: apples, oranges, ...

=, ≠

O—Ordinal

Rankings: 1st, 2nd, 3rd...

Q—Quantitative

Interval (location of zero arbitrary)

Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)

Only differences (i.e. intervals) are compared

Ratio (zero fixed)

Physical measurement: length, amounts, counts

Allow direct comparisons like twice as long

Data Scales

N—Nominal (labels or categories)

- Fruits: apples, oranges, ...

O—Ordinal

- Rankings: 1st, 2nd, 3rd...

=, ≠, **<, >**

Q—Quantitative

- **Interval** (location of zero arbitrary)
  - Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
  - Only differences (i.e. intervals) are compared
- **Ratio** (zero fixed)
  - Physical measurement: length, amounts, counts
  - Allow direct comparisons like twice as long

Data Scales

N—Nominal (labels or categories)

Fruits: apples, oranges, ...

O—Ordinal

Rankings: 1st, 2nd, 3rd...

Q—Quantitative

=, ≠, <, >, -

Interval (location of zero arbitrary)

Can measure distances or spans

Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)

Only differences (i.e. intervals) are compared

Ratio (zero fixed)

Physical measurement: length, amounts, counts

Allow direct comparisons like twice as long

Data Scales

N—Nominal (labels or categories)

Fruits: apples, oranges, ...

O—Ordinal

Rankings: 1st, 2nd, 3rd...

Q—Quantitative

Interval (location of zero arbitrary)

Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)

Only differences (i.e. intervals) are compared

=, ≠, <, >, −, **/** (%)

Ratio (zero fixed)

Can measure ratios or proportions

Physical measurement: length, amounts, counts

Allow direct comparisons like twice as long

Example

Conceptual Model

Temperature (°C)

Data Model

32.5, 54.0, -17.3, …

Decimal numbers

Data Scales

Temperature Value (Q)

Burned vs. Not-Burned (N) — Derived

Hot, Warm, Cold (O) — Derived

Dimensions & Measures

Dimensions (~ independent variables)

Often discrete variables describing data (N, O)

Categories, dates, binned quantities

Measures (~ dependent variables)

Continuous values that can be aggregated (Q)

Numbers to be analyzed

Aggregate as sum, count, average, std. dev...

Not a strict distinction. The same variable may be treated either way depending on the task (e.g. Year: 2001, 2002 ...).

Example: U.S. Census Data

U.S. Census Data

Year: 1850 – 2000 (every decade)

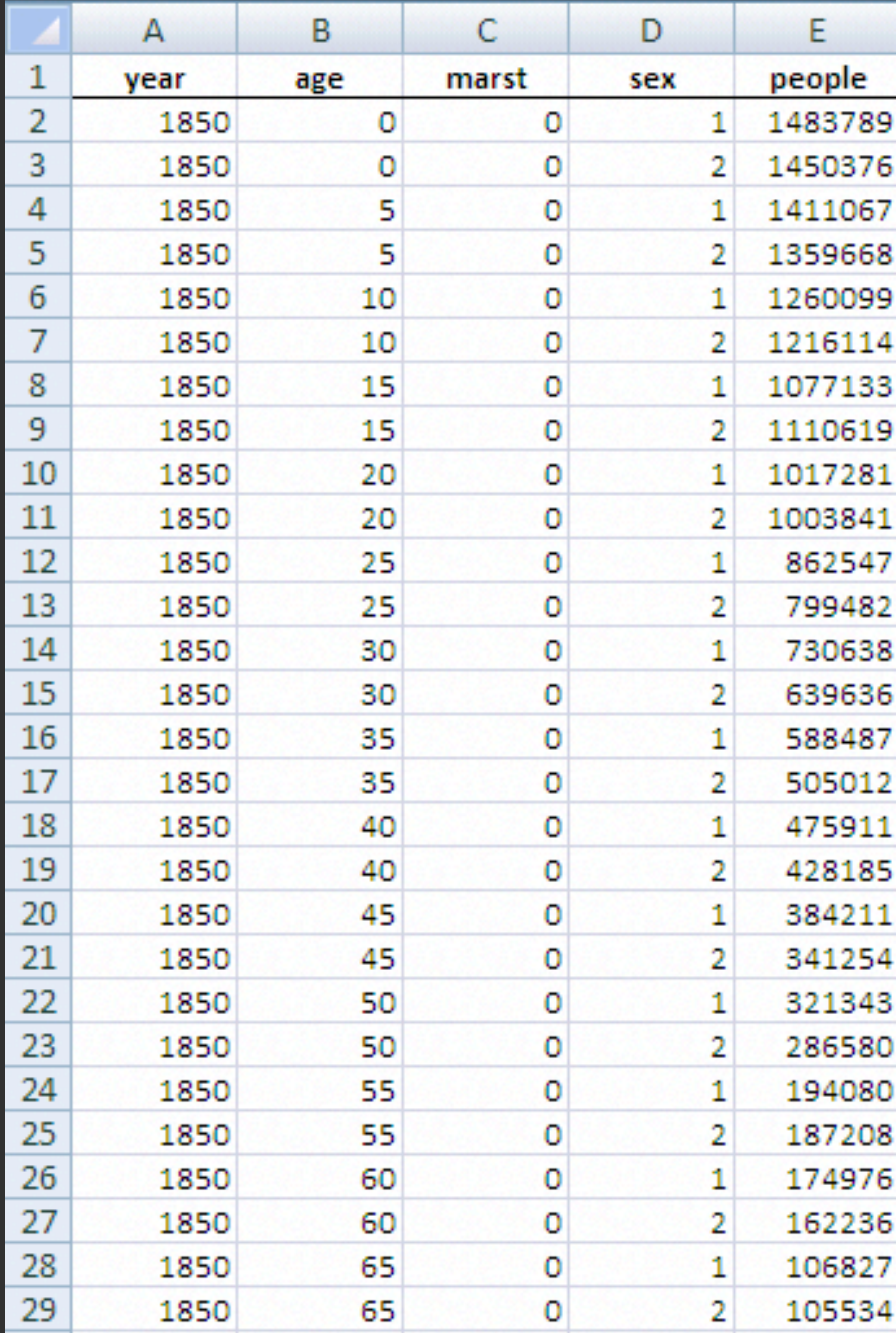
Age: 0 – 90+

Marital Status: Single, Married, Divorced,

Sex: Male, Female

People Count: # of people in group

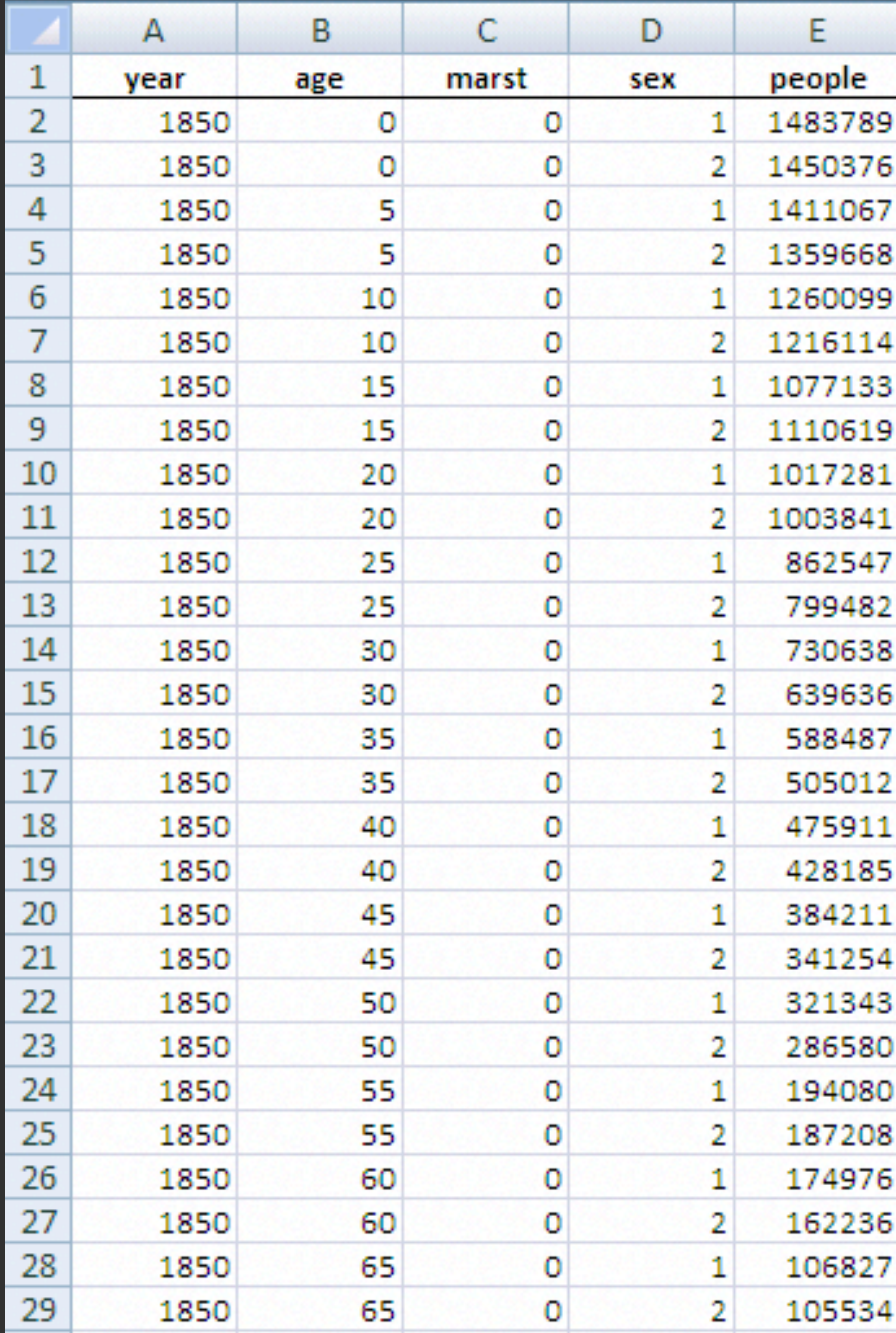
2,348 data points



	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080
25	1850	55	0	2	187208
26	1850	60	0	1	174976
27	1850	60	0	2	162236
28	1850	65	0	1	106827
29	1850	65	0	2	105534

U.S. Census Data

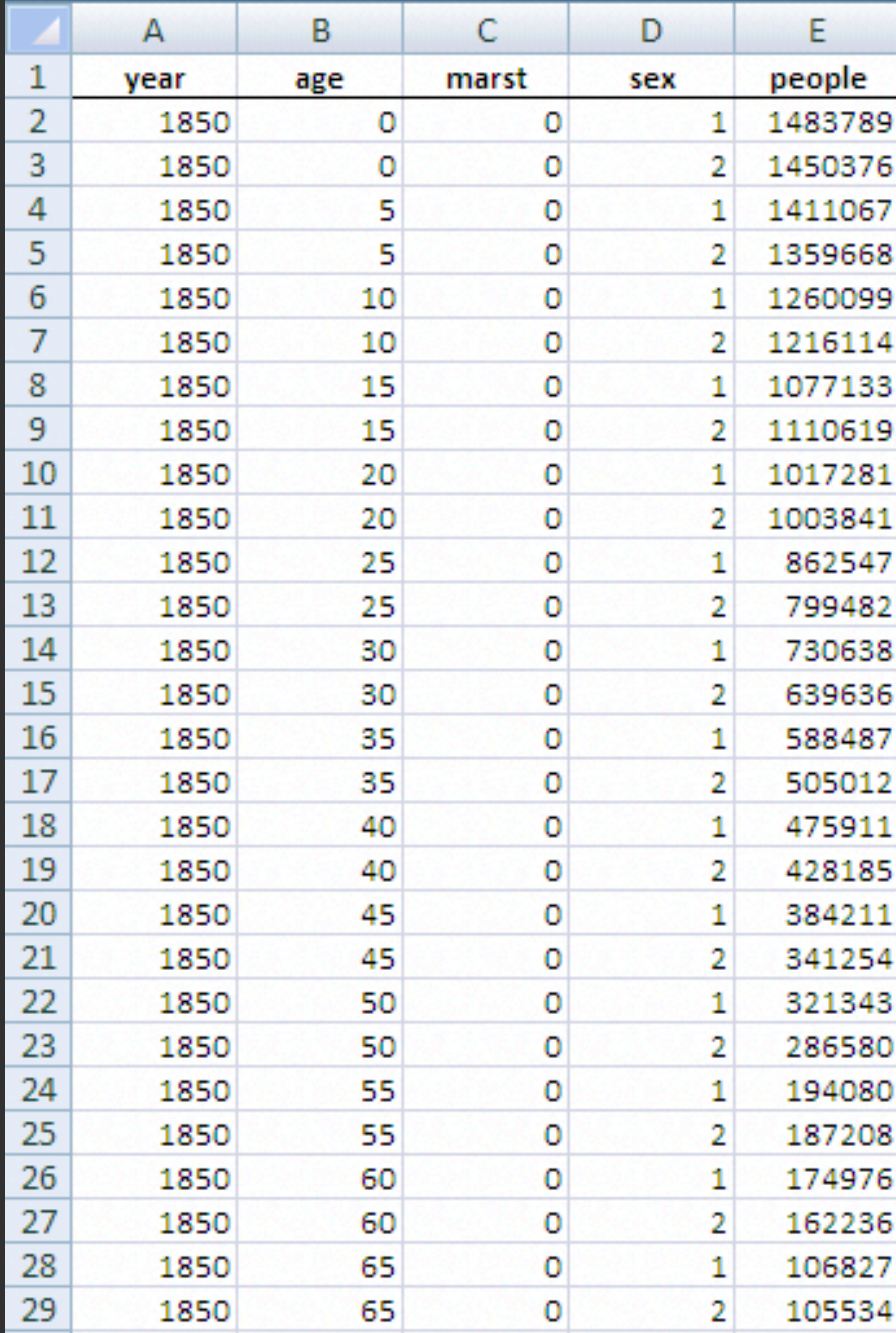
Year	Q-Interval (O)
Age	Q-Ratio (O)
Marital Status	N
Sex	N
People Count	Q-Ratio



	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080
25	1850	55	0	2	187208
26	1850	60	0	1	174976
27	1850	60	0	2	162236
28	1850	65	0	1	106827
29	1850	65	0	2	105534

U.S. Census Data

Year	Depends!
Age	Depends!
Marital Status	Dimension
Sex	Dimension
People Count	Measure



	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080
25	1850	55	0	2	187208
26	1850	60	0	1	174976
27	1850	60	0	2	162236
28	1850	65	0	1	106827
29	1850	65	0	2	105534

Image Models

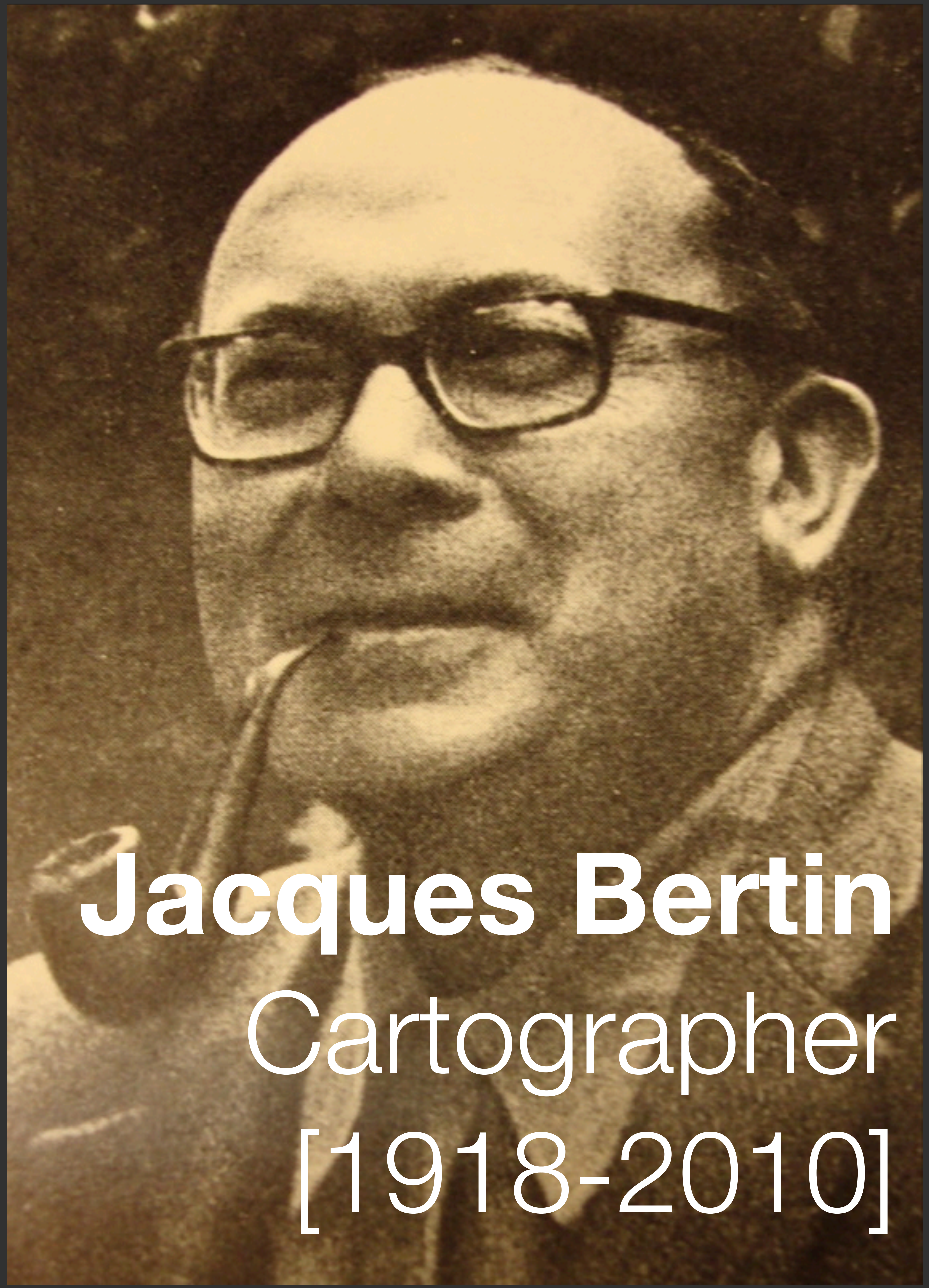
Visual Language is a Sign System

Images perceived as a set of signs

Sender encodes information in signs

Receiver decodes information from sign

Semiology of Graphics, 1967



Jacques Bertin
Cartographer
[1918-2010]

Image Models

Visual Marks

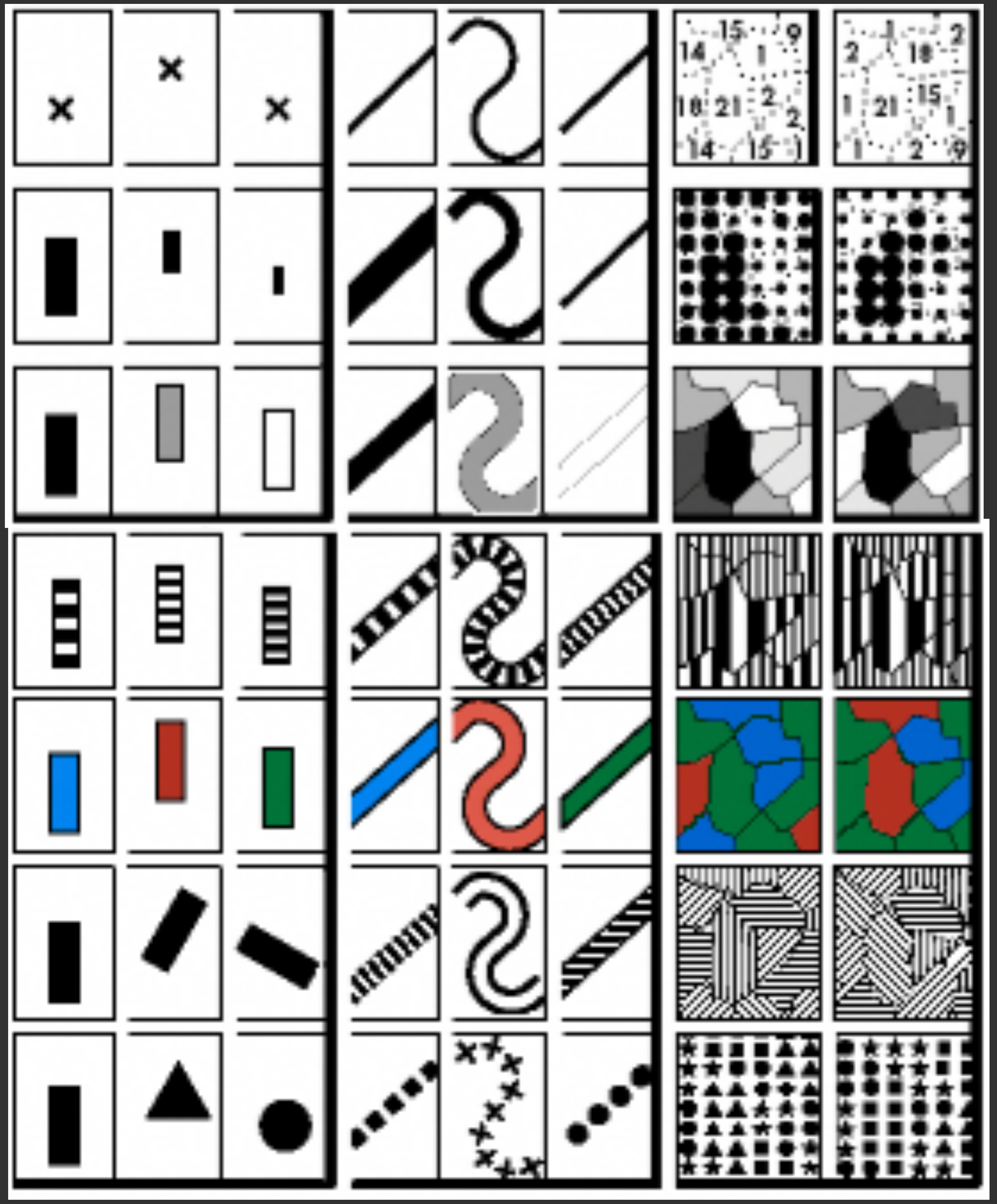
Basic graphical elements in an image

Represent information

Perceptual Channels

Control the appearance of marks

Encode information



Points Lines Areas

Position

Size

Value

Texture

Color

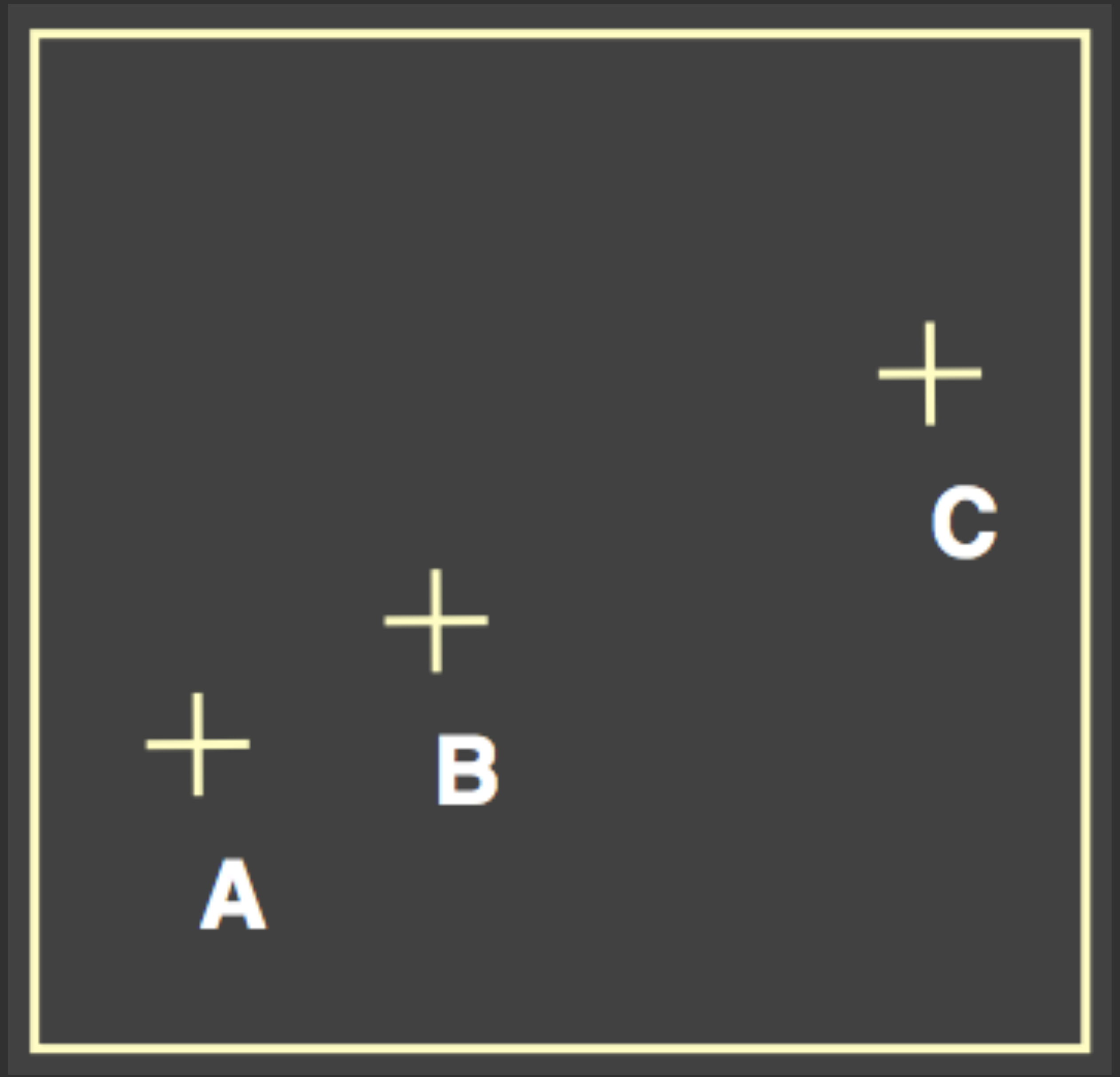
Orientation

Shape

Coding Information in Position

1. A, B, C are distinguishable
2. B is between A and C.
3. BC is twice as long as AB.

∴ Encode quantitative variables (Q)



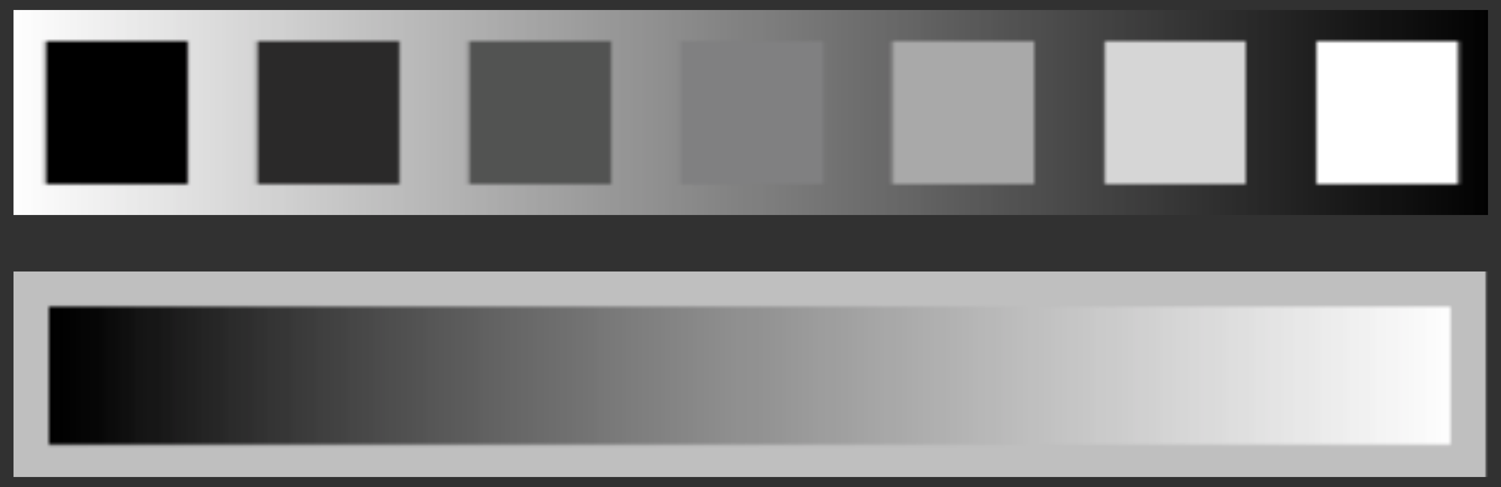
"Resemblance, order and proportional are the three signfields in graphics." — Bertin

Coding Information in Color and Value

Value (lightness) is perceived as ordered

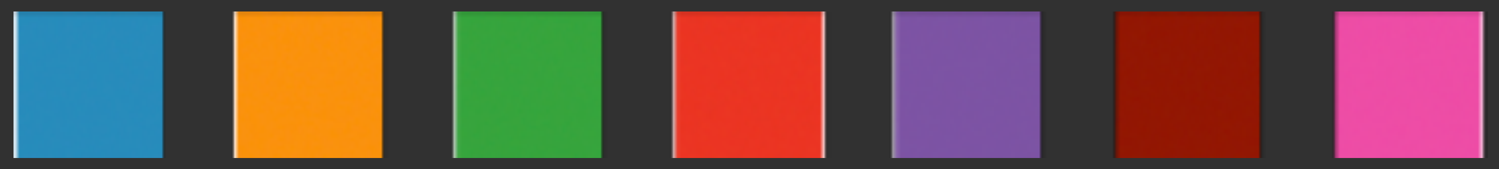
∴ Encode ordinal variables (O) *[better]*

∴ Encode continuous variables (Q)



Hue is normally perceived as unordered

∴ Encode nominal variables (N)



Bertin’s Levels of Organization

Position	N	O	Q
Size	N	O	Q
Value	N	O	Q
Texture	N	o	
Color	N		
Orientatio	N		
Shape	N		

Nominal

Ordinal

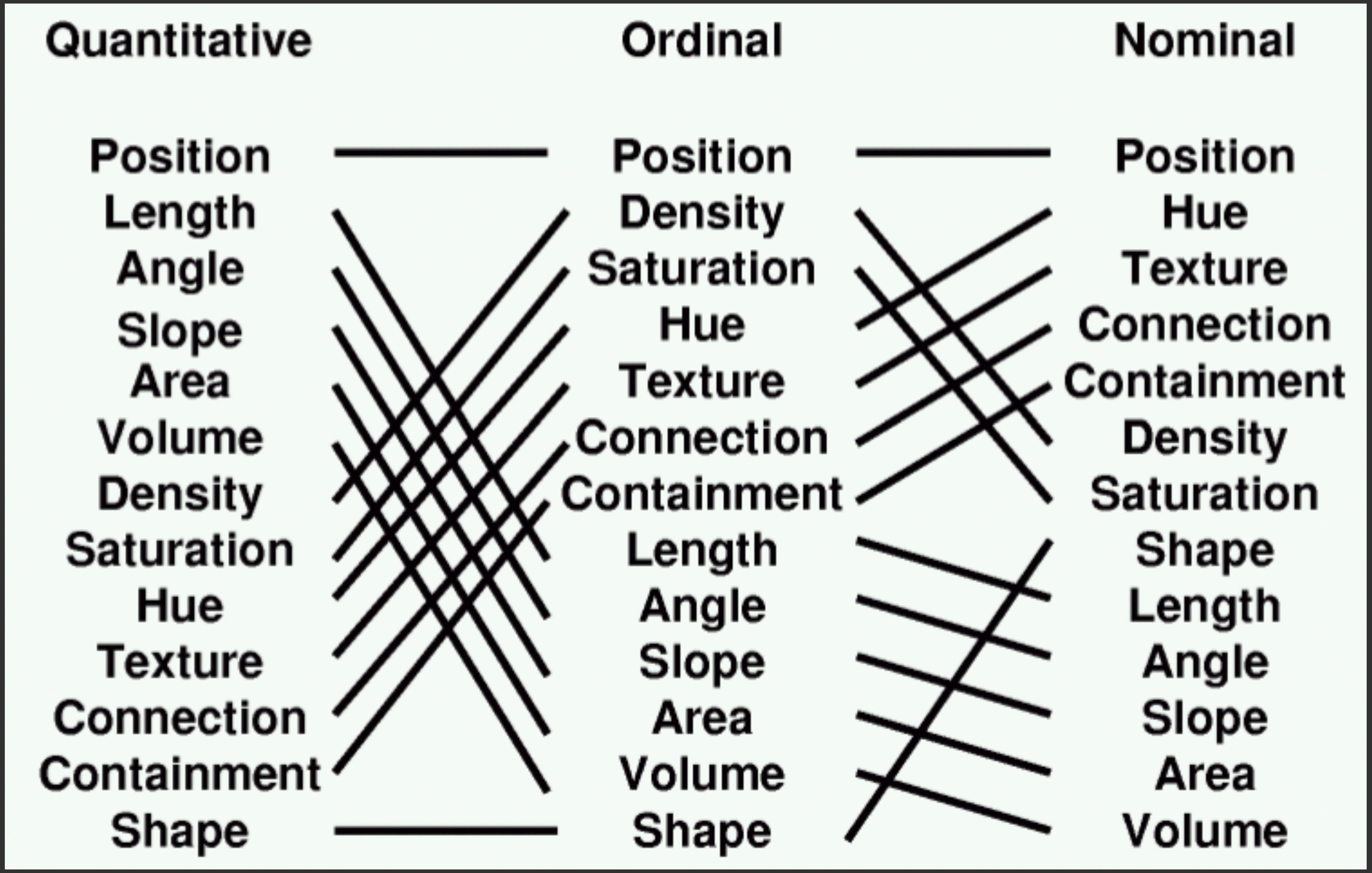
Quantitative

Note: **Q** ⊂ **O** ⊂ **N**

Mackinlay’s Ranking

Expanded Bertin’s variables and conjectured effectiveness of encodings by data type.

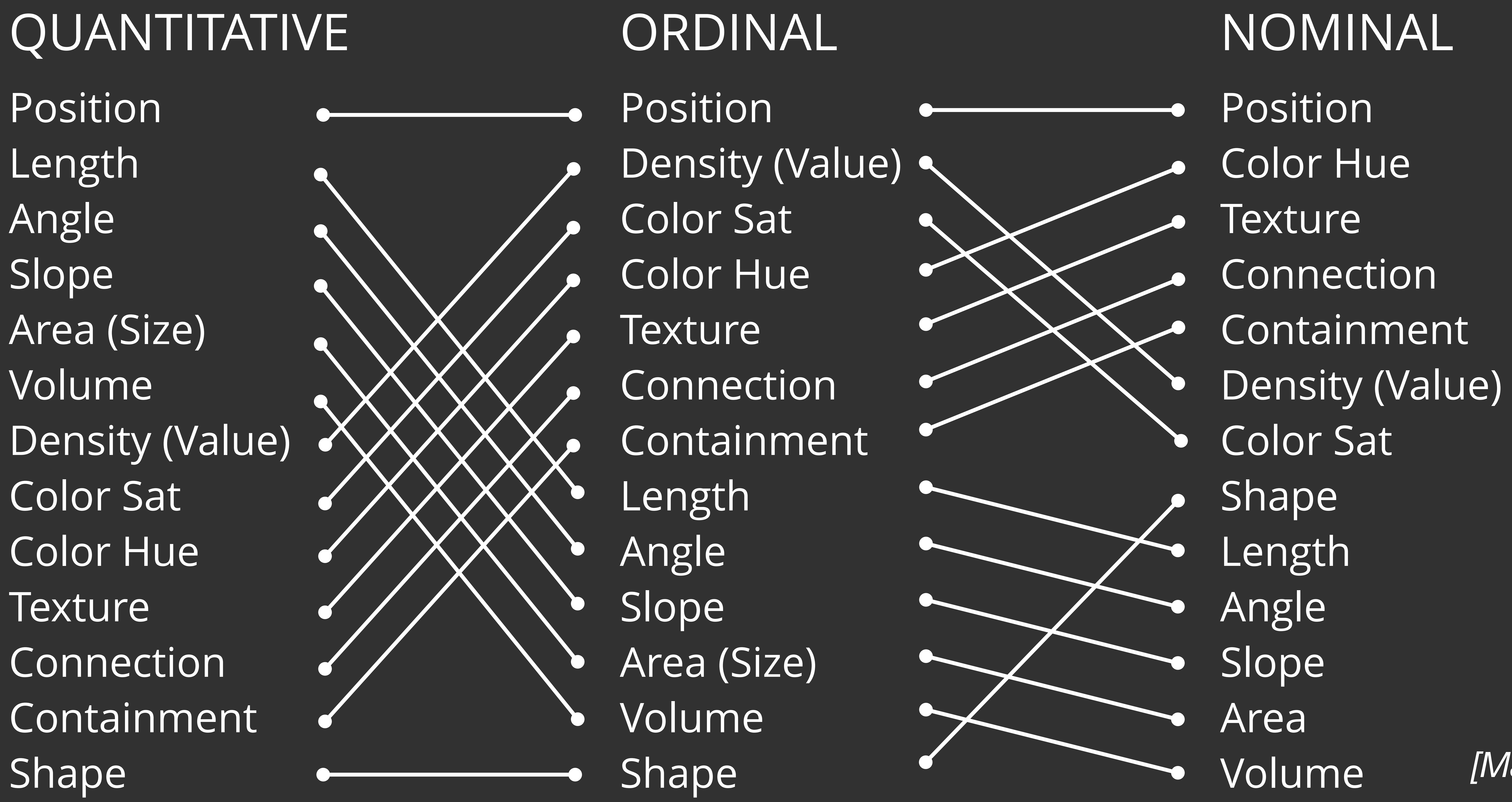
[Mackinlay 86]



Quantitative	Ordinal	Nominal
Position	Position	Position
Length	Density	Hue
Angle	Saturation	Texture
Slope	Hue	Connection
Area	Texture	Containment
Volume	Connection	Density
Density	Containment	Saturation
Saturation	Length	Shape
Hue	Angle	Length
Texture	Slope	Angle
Connection	Area	Slope
Containment	Volume	Area
Shape	Shape	Volume

Jock D. Mackinlay
Vice President
Tableau Software

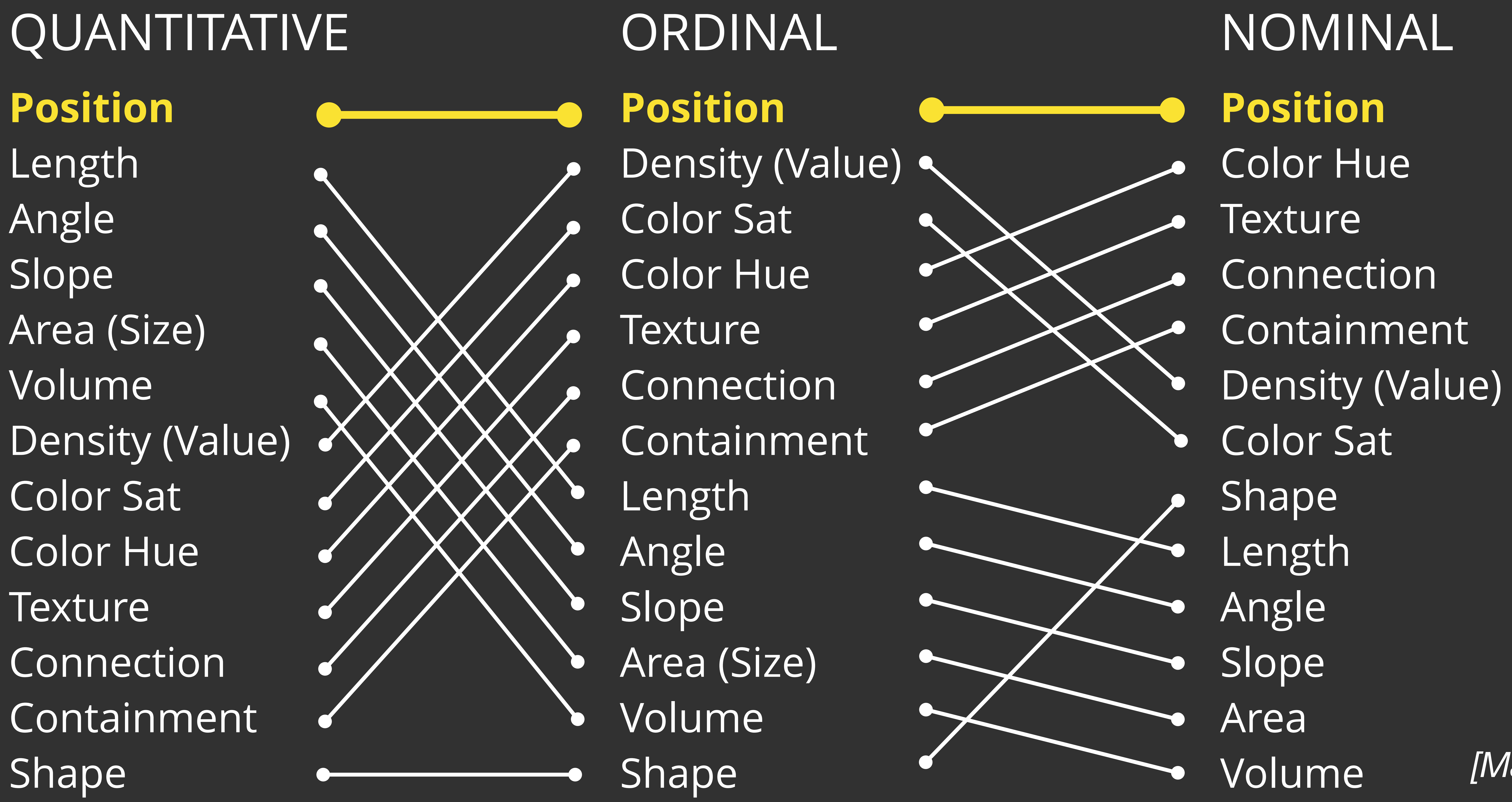
Effectiveness Rankings



QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density (Value)	Color Hue
Angle	Color Sat	Texture
Slope	Color Hue	Connection
Area (Size)	Texture	Containment
Volume	Connection	Density (Value)
Density (Value)	Containment	Color Sat
Color Sat	Length	Shape
Color Hue	Angle	Length
Texture	Slope	Angle
Connection	Area (Size)	Slope
Containment	Volume	Area
Shape	Shape	Volume

[Mackinlay 86]

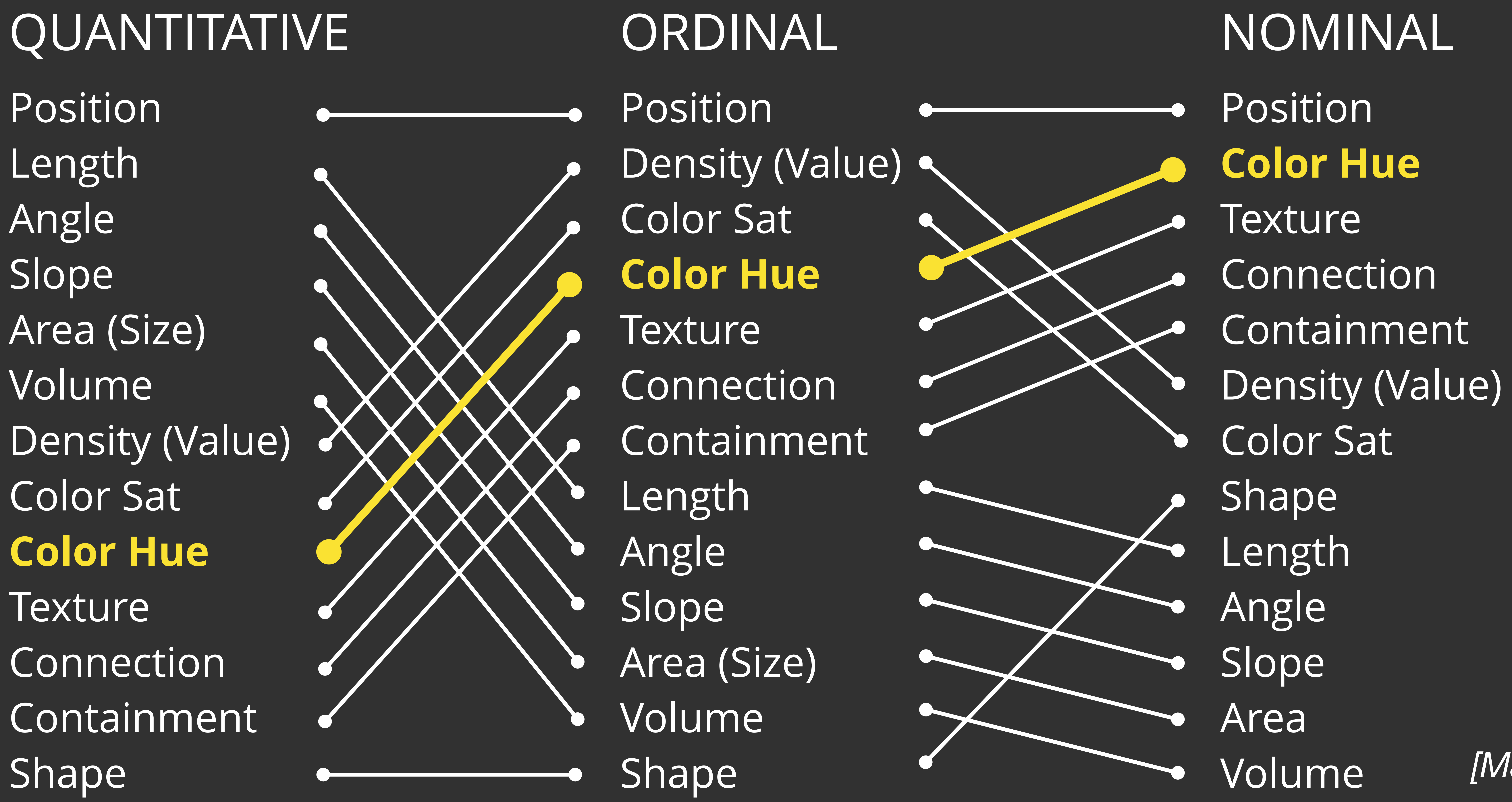
Effectiveness Rankings



QUANTITATIVE	ORDINAL	NOMINAL
Position	**Position**	**Position**
Length	Density (Value)	Color Hue
Angle	Color Sat	Texture
Slope	Color Hue	Connection
Area (Size)	Texture	Containment
Volume	Connection	Density (Value)
Density (Value)	Containment	Color Sat
Color Sat	Length	Shape
Color Hue	Angle	Length
Texture	Slope	Angle
Connection	Area (Size)	Slope
Containment	Volume	Area
Shape	Shape	Volume

[Mackinlay 86]

Effectiveness Rankings

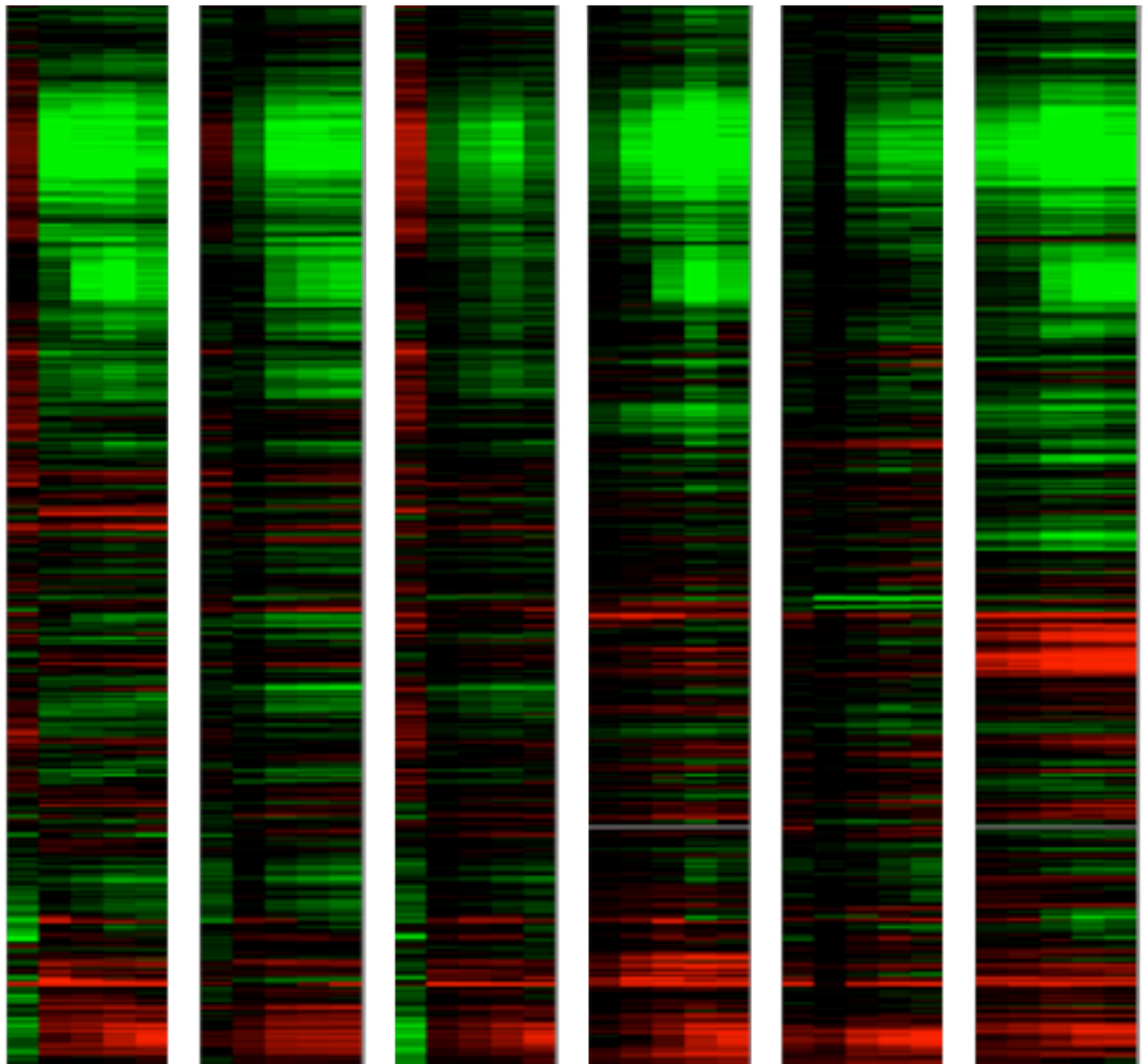


QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density (Value)	**Color Hue**
Angle	Color Sat	Texture
Slope	**Color Hue**	Connection
Area (Size)	Texture	Containment
Volume	Connection	Density (Value)
Density (Value)	Containment	Color Sat
Color Sat	Length	Shape
Color Hue	Angle	Length
Texture	Slope	Angle
Connection	Area (Size)	Slope
Containment	Volume	Area
Shape	Shape	Volume

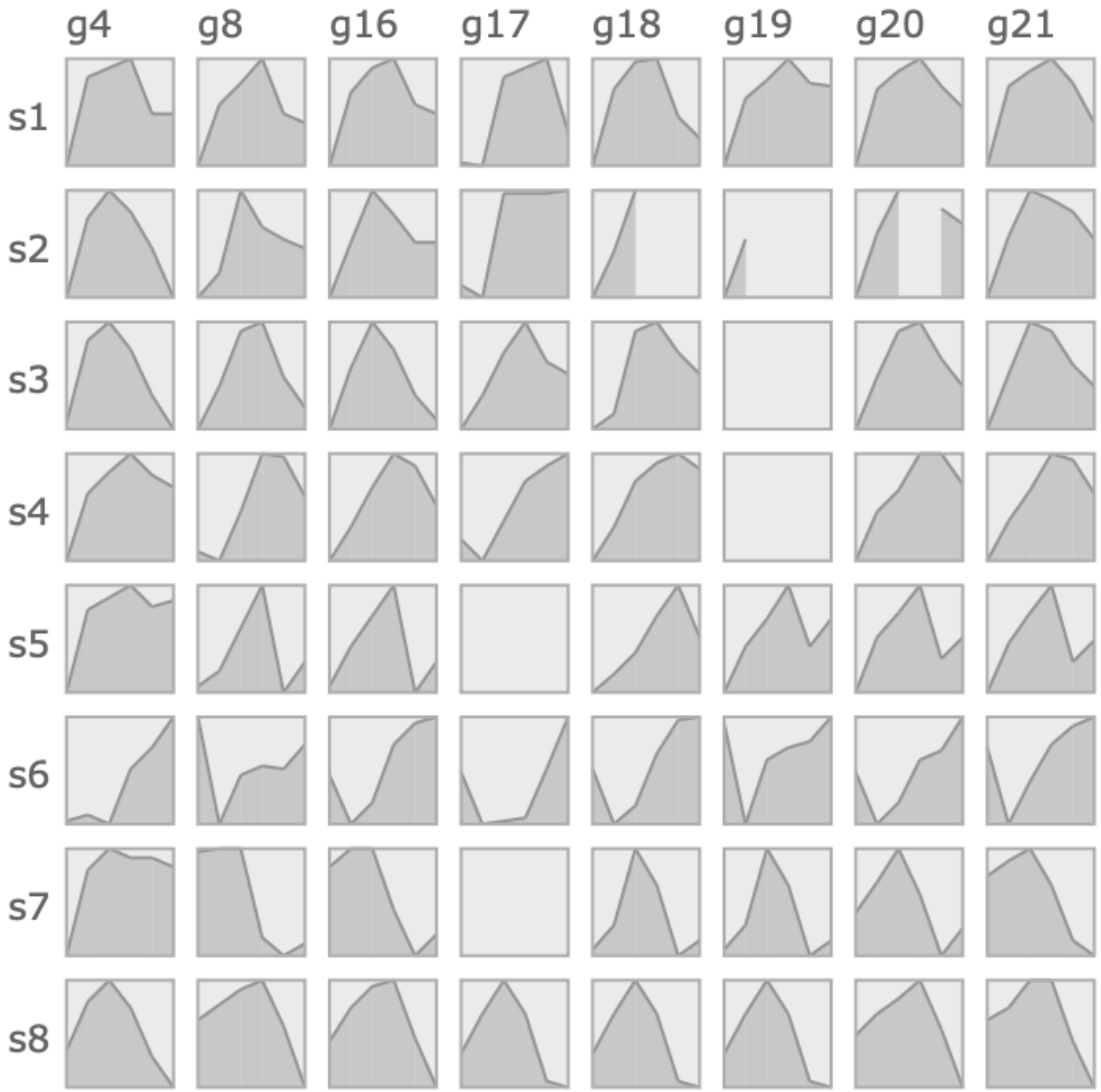
[Mackinlay 86]

Gene Expression Time-Series [Meyer et al ’11]

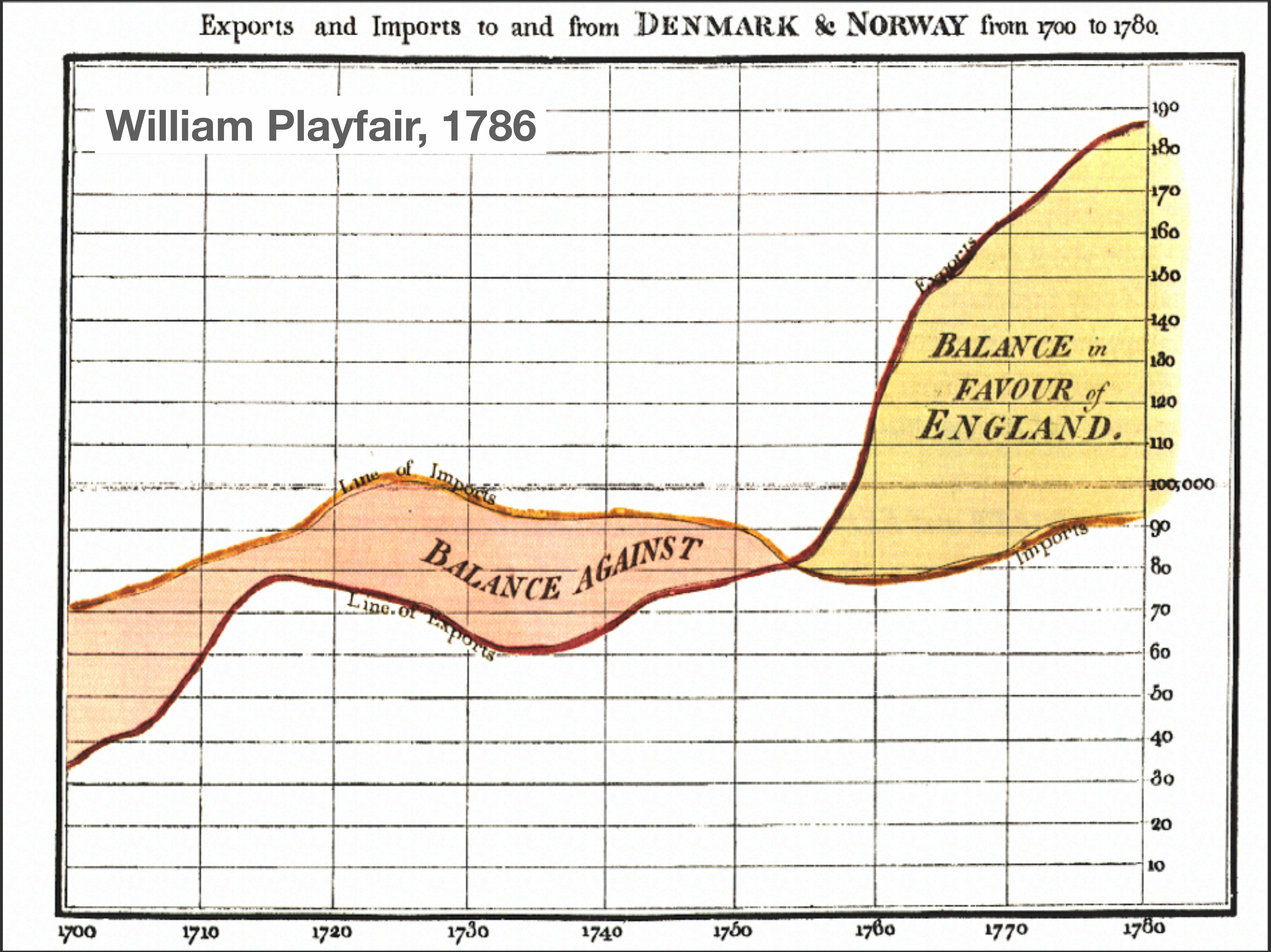
Heatmap



Curvemap

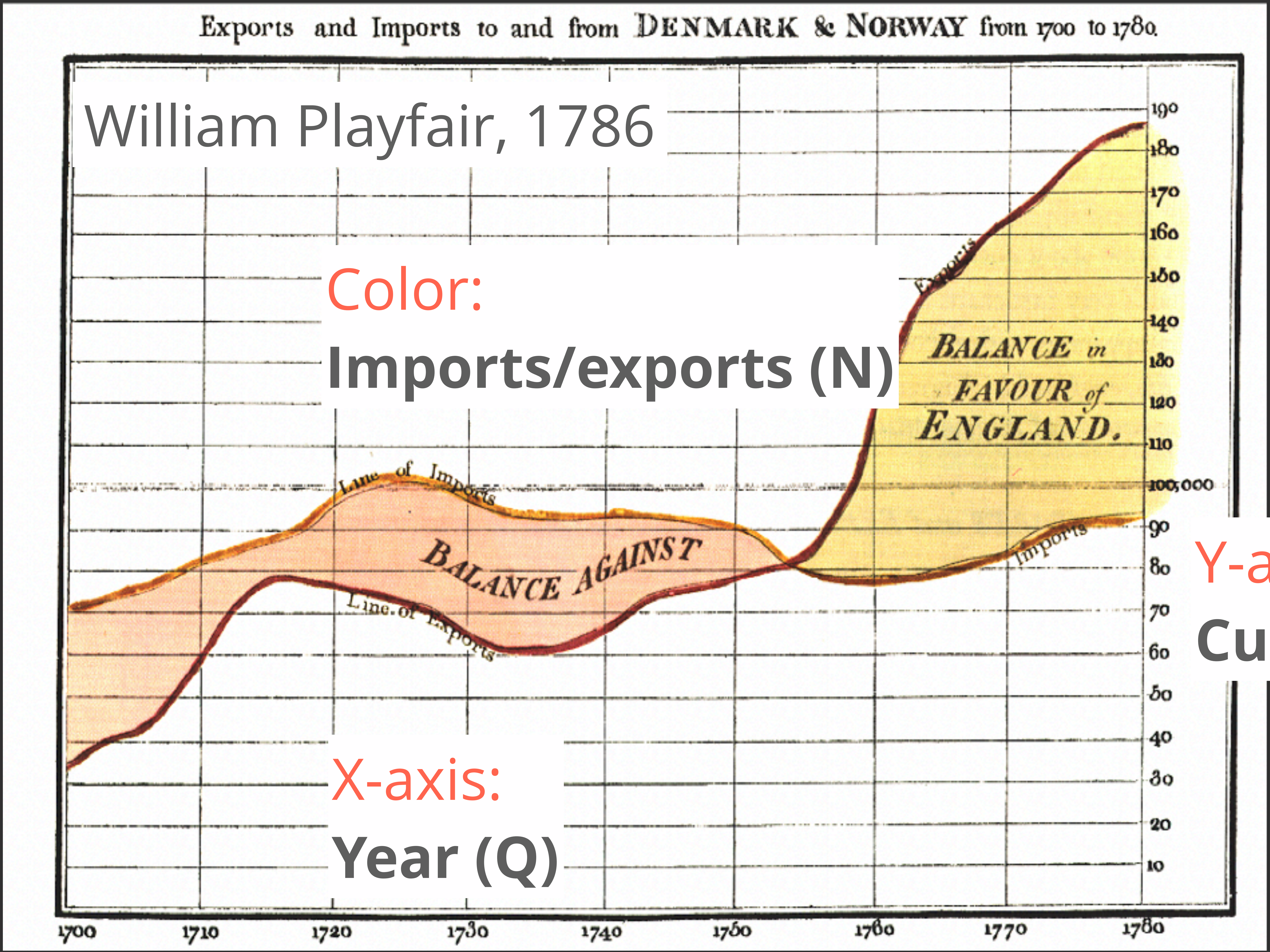


Example: Deconstructions



Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.

William Playfair, 1786



William Playfair, 1786

Color:

Imports/exports (N)

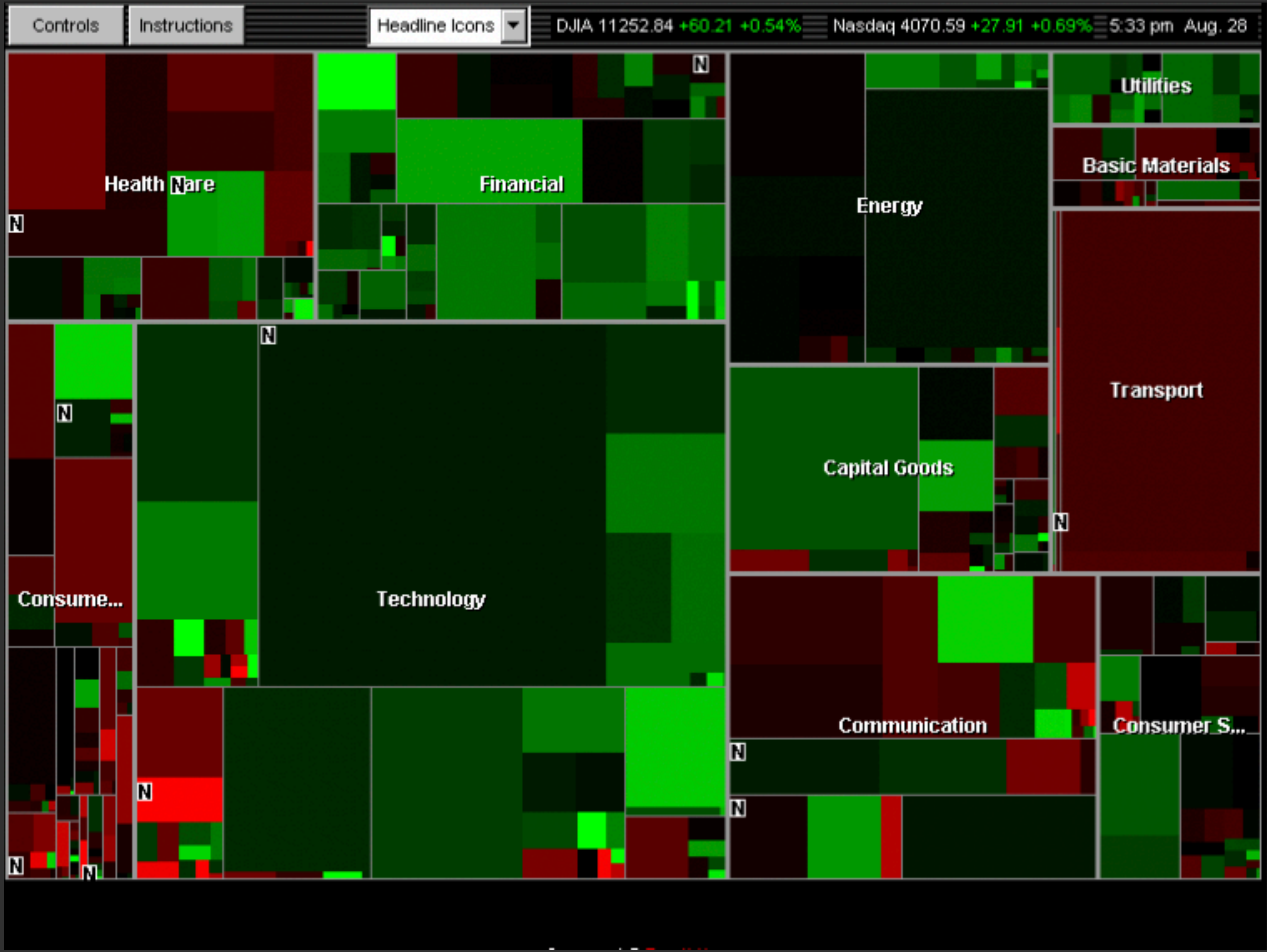
X-axis:

Year (Q)

Y-axis:

Currency (Q)

Wattenberg's Map of the Market

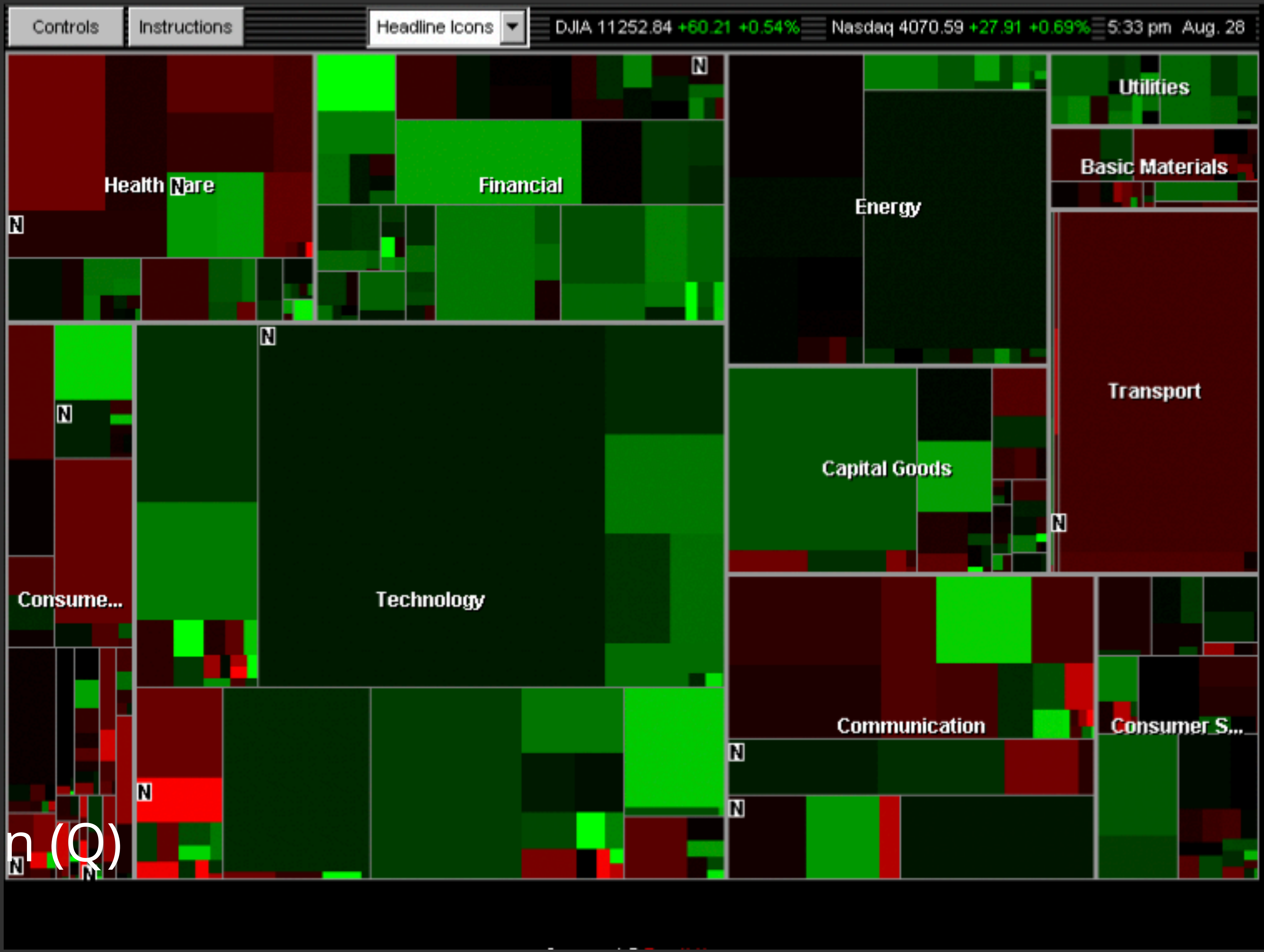


Rectangle Area:
market cap (Q)

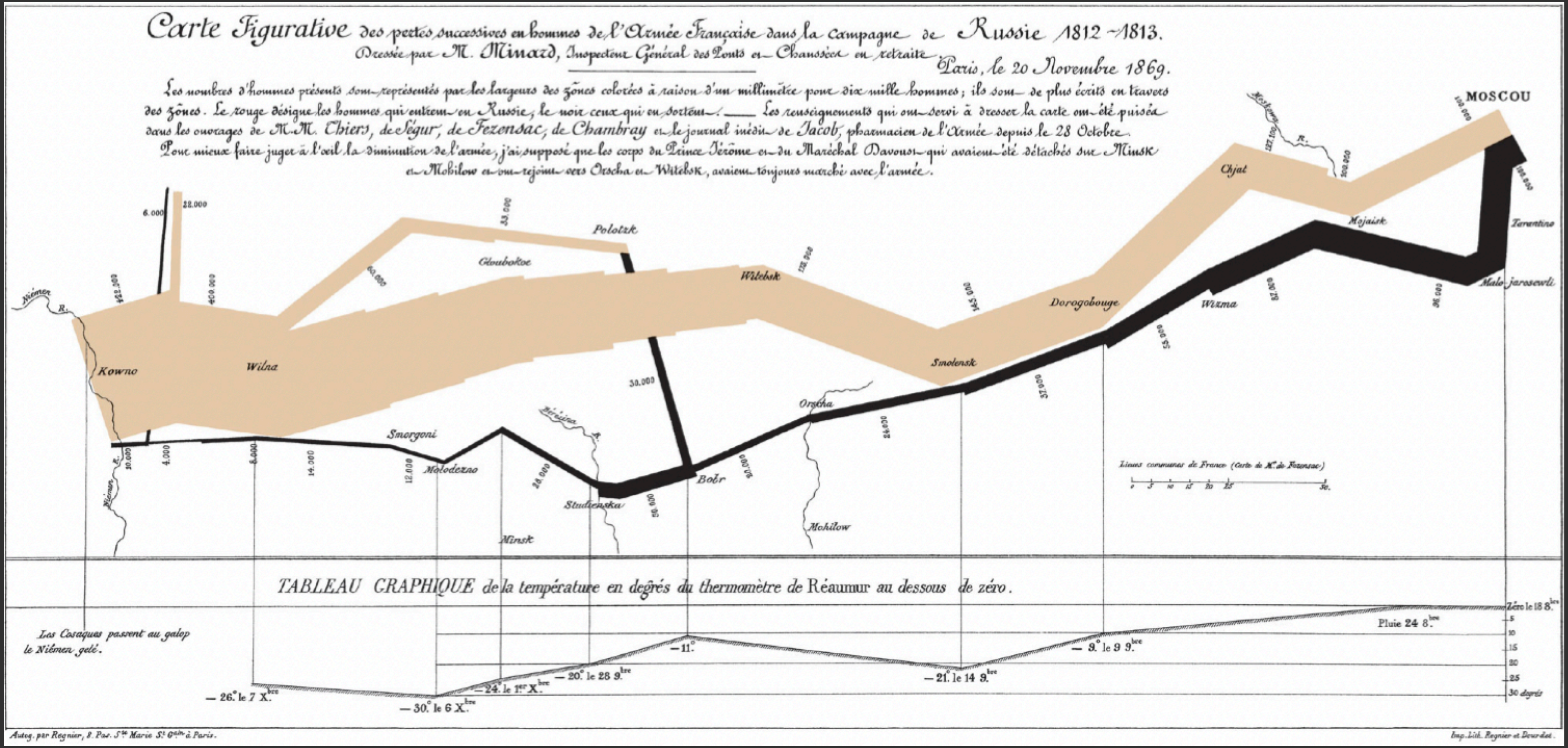
Rectangle Position:
market sector (N),
market cap (Q)

Color Hue:
loss vs. gain (N)

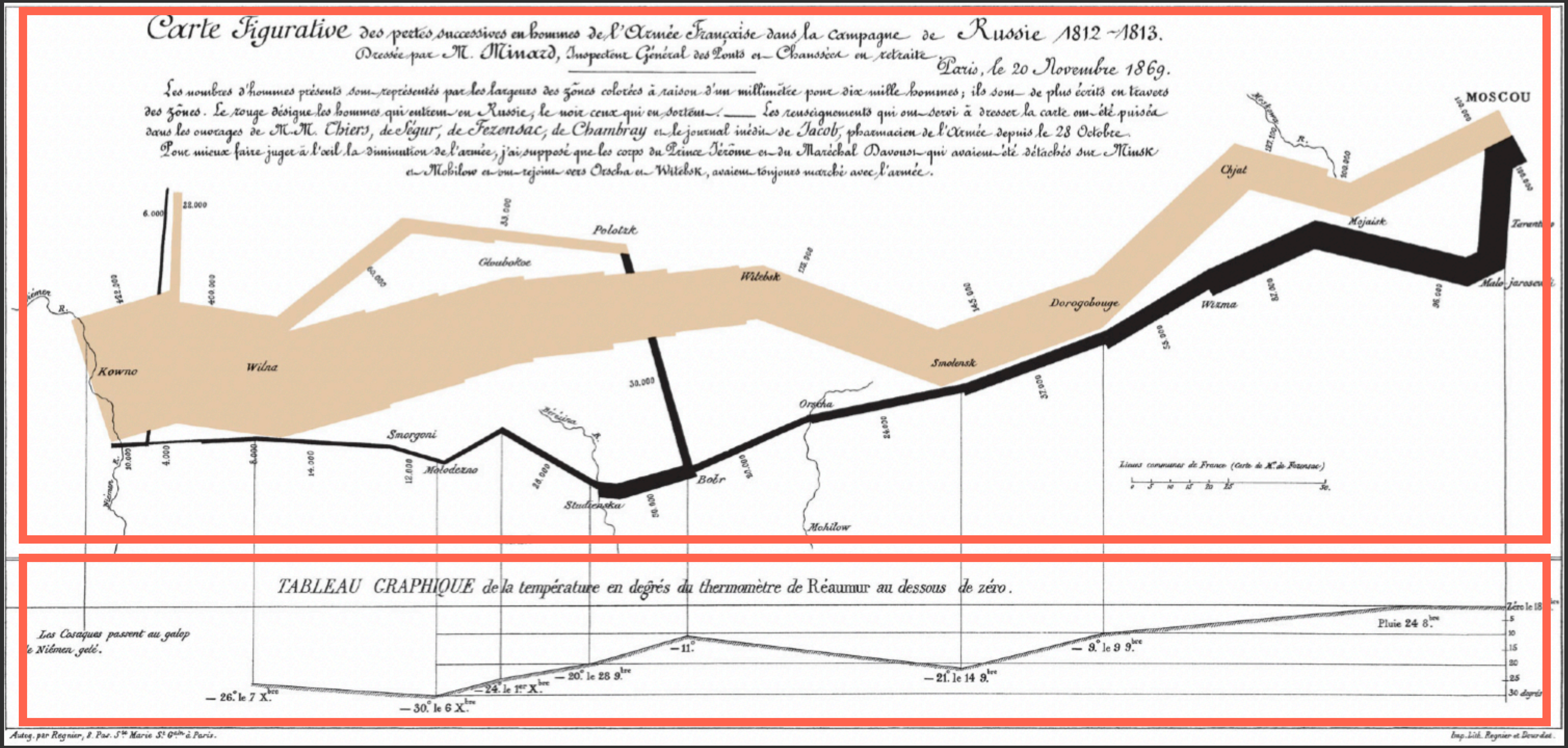
Color Value:
magnitude of loss or gain (Q)



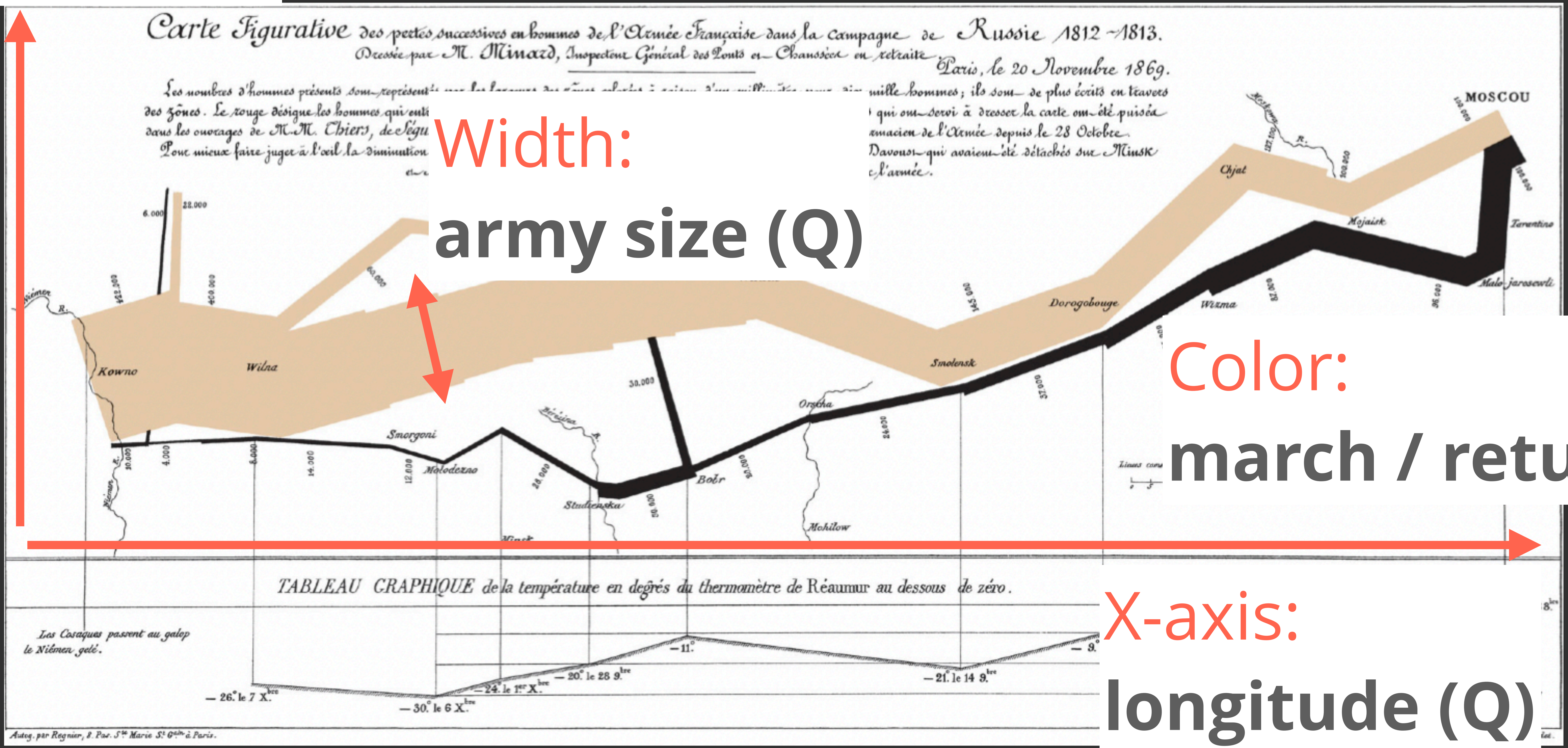
Minard 1869: Napoleon’s March



Minard 1869: Napoleon’s March



Minard 1869: Napoleon's March



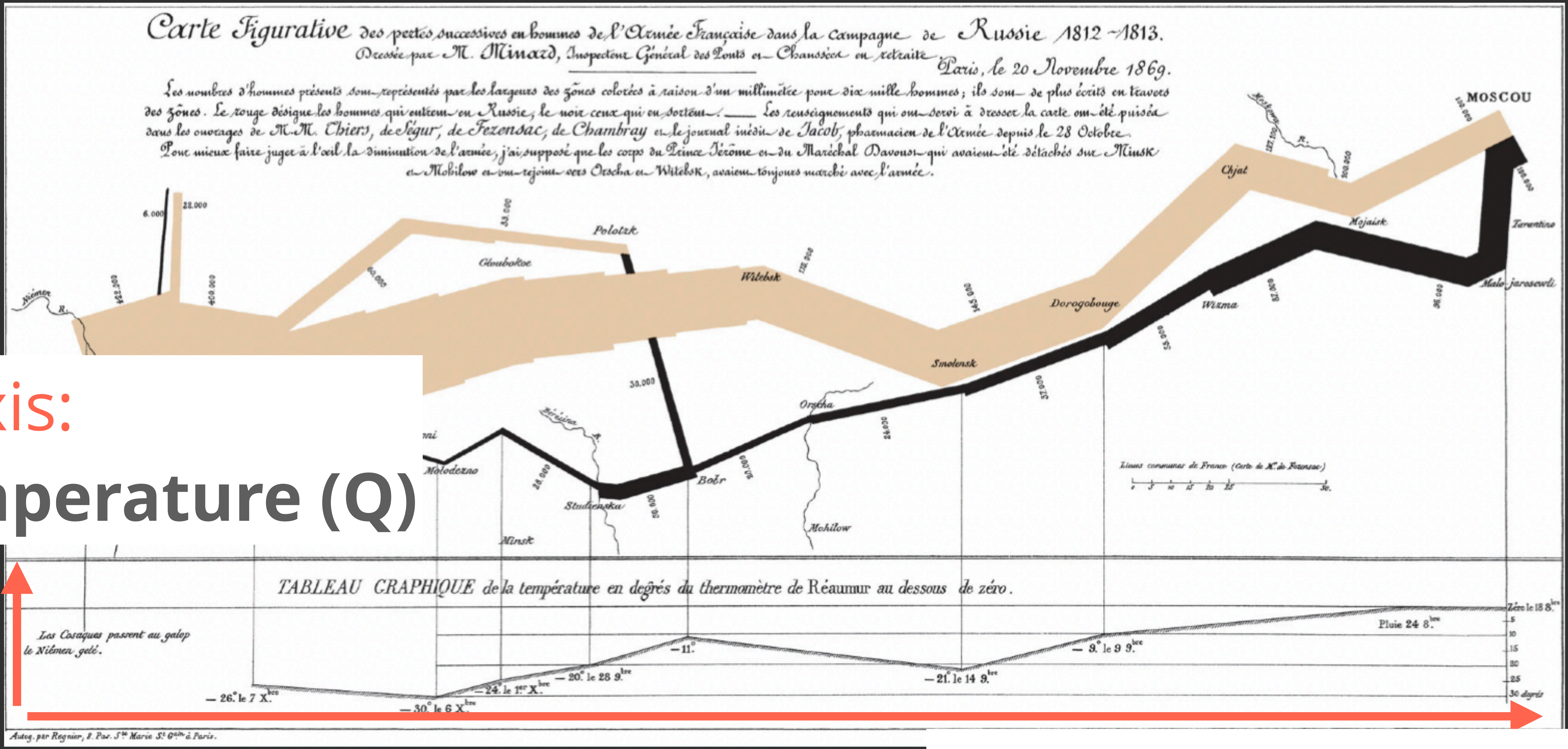
Y-axis:
latitude (Q)

Width:
army size (Q)

Color:
march / return

X-axis:
longitude (Q)

Minard 1869: Napoleon’s March



Y-axis:

temperature (Q)

X-axis:

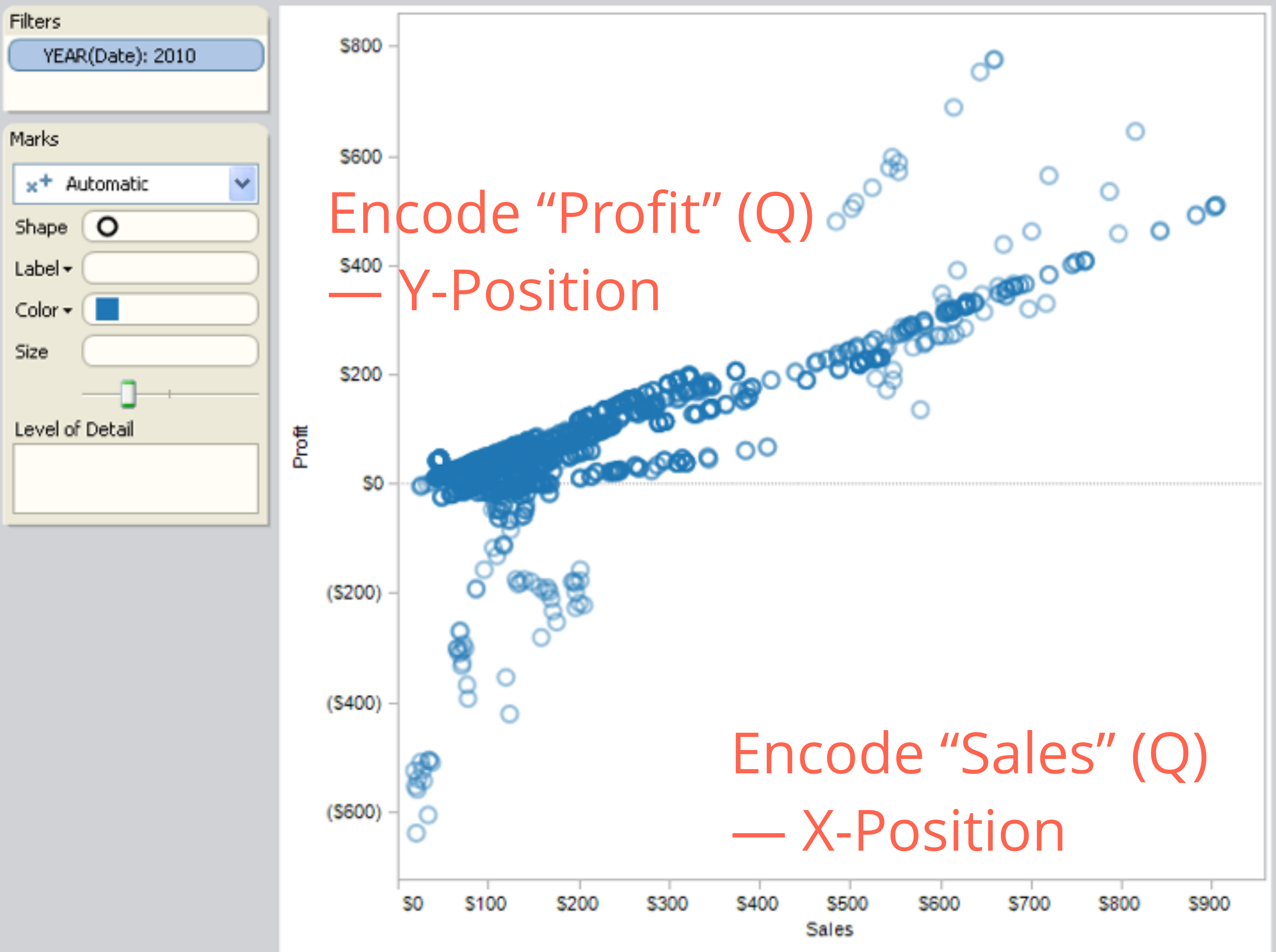
longitude (Q) / time (O)

Example: Encoding Data

Example: Coffee Sales

Sales figures for a fictional coffee chain

Sales	Q-Ratio
Profit	Q-Ratio
Marketing	Q-Ratio
Product Type	N {Coffee, Espresso, Herbal Tea, Tea}
Market	N {Central, East, South, West}



Filters

YEAR(Date): 2010

Marks

Automatic

Shape ○

Label

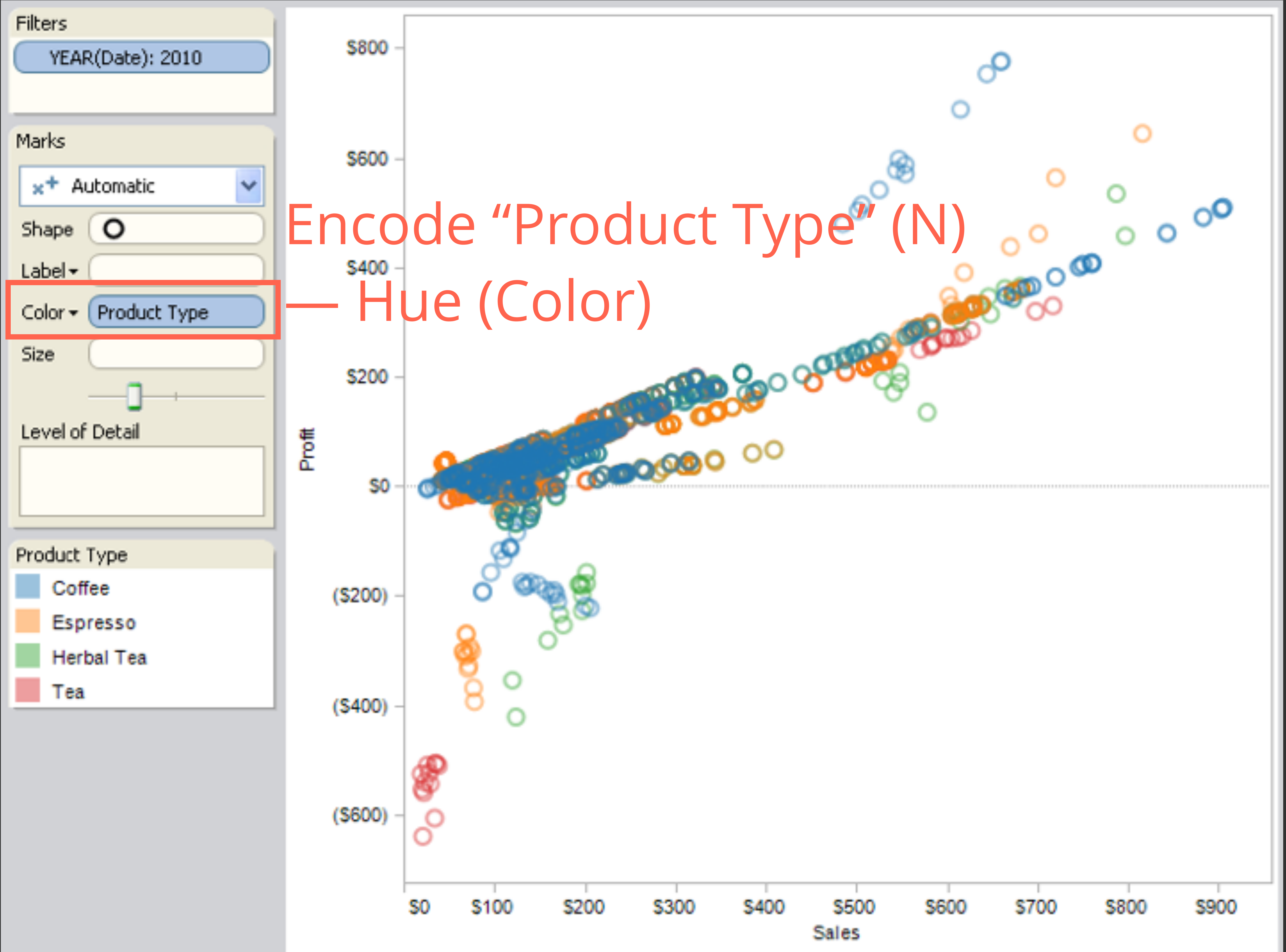
Color 🟦

Size

Level of Detail

Encode “Profit” (Q) — Y-Position

Encode “Sales” (Q) — X-Position



Filters

YEAR(Date): 2010

Marks

Automatic

Shape ○

Label

Color Product Type

Size

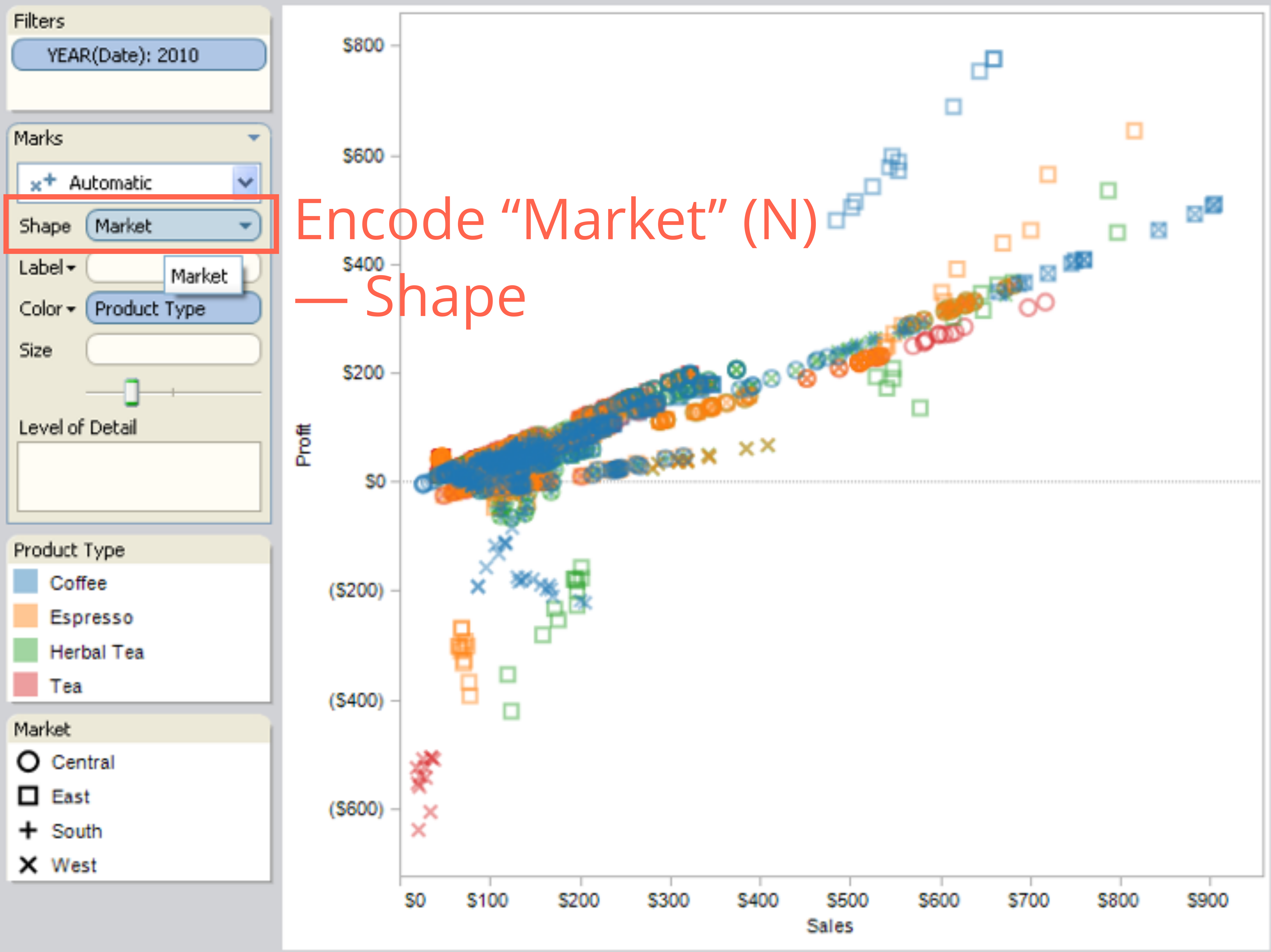
Level of Detail

Product Type

- 🟦 Coffee
- 🟧 Espresso
- 🟩 Herbal Tea
- 🟥 Tea

Encode “Product Type” (N)

— Hue (Color)



Filters

YEAR(Date): 2010

Marks

Automatic

Shape: Market

Label

Color: Product Type

Size

Level of Detail

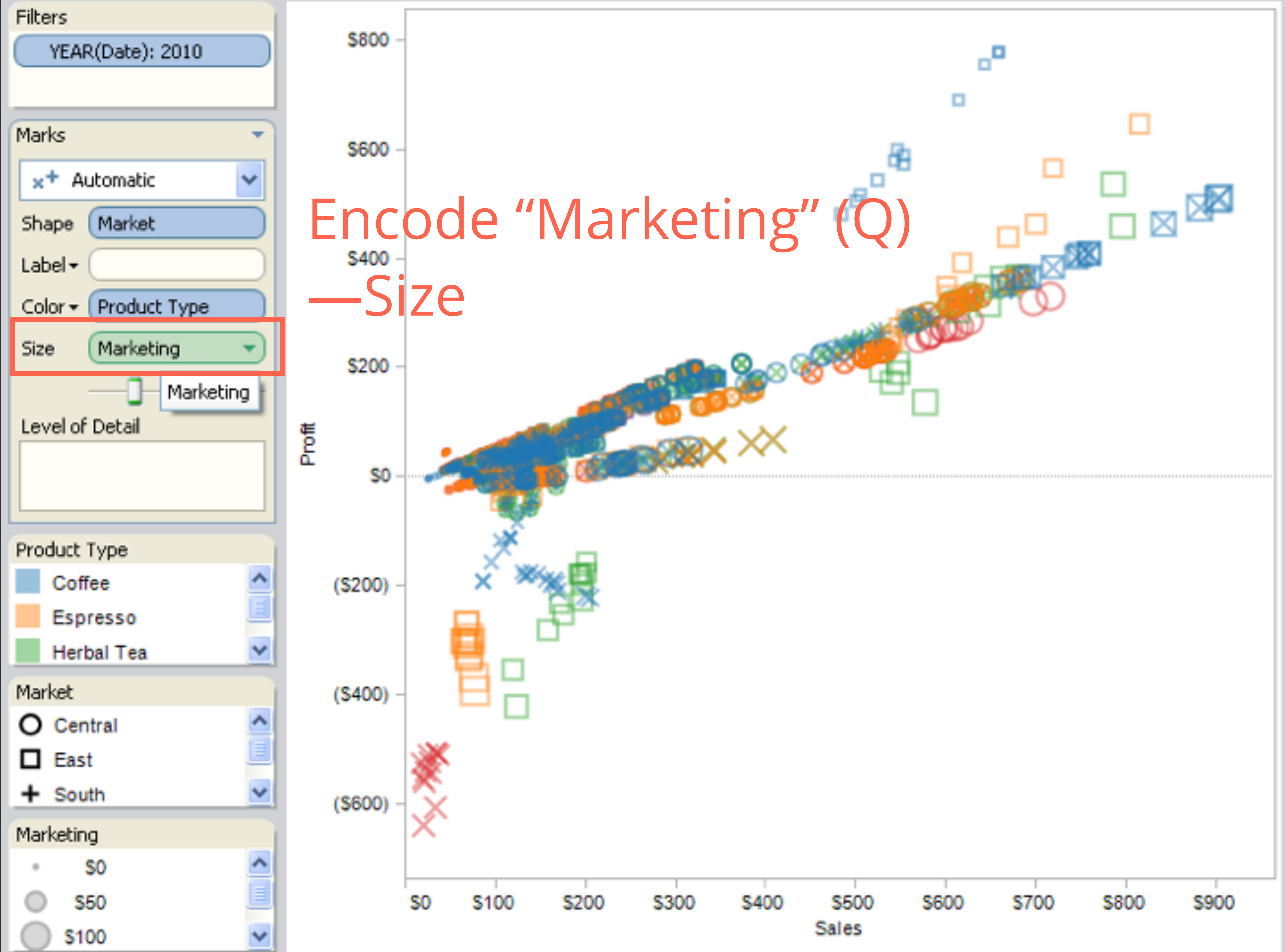
Product Type

- 🟦 Coffee
- 🟧 Espresso
- 🟩 Herbal Tea
- 🟥 Tea

Market

- ○ Central
- □ East
- + South
- × West

Encode “Market” (N)
— Shape



Encode “Marketing” (Q) —Size

Filters

YEAR(Date): 2010

Marks

Automatic

Shape: Market

Label:

Color: Product Type

Size: Marketing

Level of Detail

Product Type

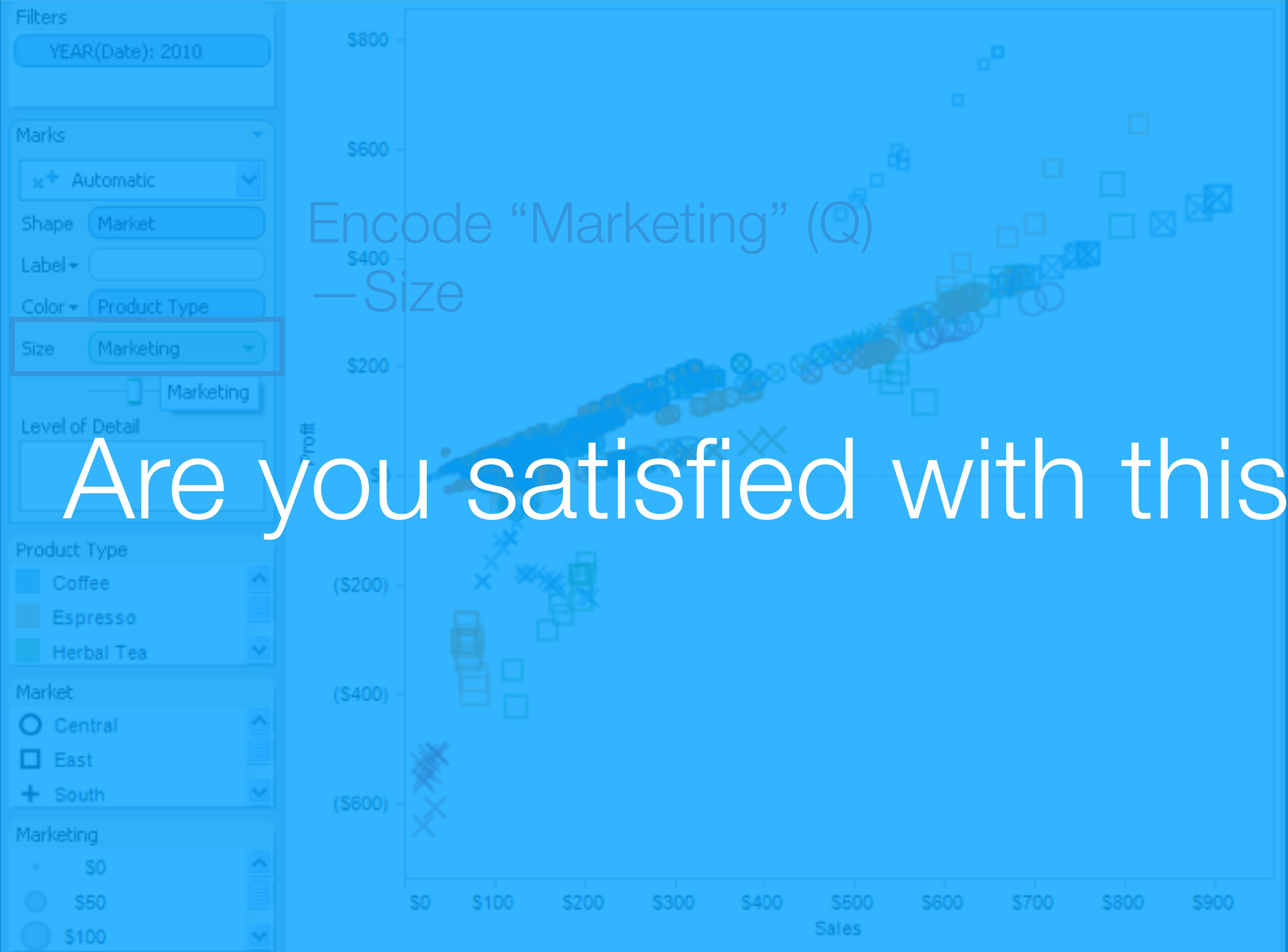
- 🟦 Coffee
- 🟧 Espresso
- 🟩 Herbal Tea

Market

- ○ Central
- □ East
- + South

Marketing

- • $0
- ● $50
- ● $100



Filters

YEAR(Date): 2010

Marks

Automatic

Shape: Market

Label

Color: Product Type

Size: Marketing

Marketing

Level of Detail

Product Type

- 🟦 Coffee
- 🟫 Espresso
- 🟩 Herbal Tea

Market

- ○ Central
- □ East
- + South

Marketing

- $0
- $50
- $100

Encode “Marketing” (Q) —Size

Are you satisfied with this chart?

Avoid over-encoding

Use trellis plots (small multiples/facets) that subdivide space to enable comparison across multiple plots.

Market
Central
East
South
West
Profit
$500
$0
($500)
$0
$200
$400
$600
$800
Sales

Formalizing Design

Choosing visual encodings

Assume **k** visual channels and **n** data attributes. We would like to pick the “best” encoding among a combinatorial set of possibilities of size $\mathbf{(n+1)^k}$

Choosing visual encodings

Assume k visual encodings and n data attributes. We would like to pick the “best” encoding among a combinatorial set of possibilities of size $(n+1)^k$

Principle of Consistency

The properties of the image (visual variables) should match the properties of the data.

Principle of Importance Ordering

Encode the most important information in the most effective way.

Design Criteria [Mackinlay 86]

Expressiveness

Effectiveness

Design Criteria

Expressiveness

A set of facts is expressible in a visualization if it expresses all the facts and only the facts in the data.

Effectiveness

[Mackinlay 86]

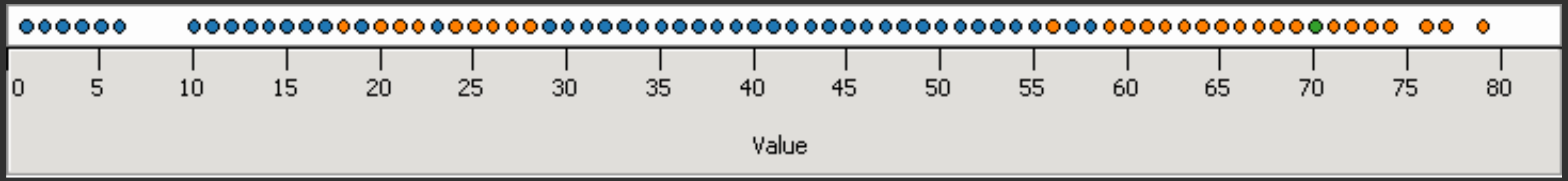
Design Criteria Translated

Tell the truth and nothing but the truth

(don’t lie, and don’t lie by omission)

Can not express the facts

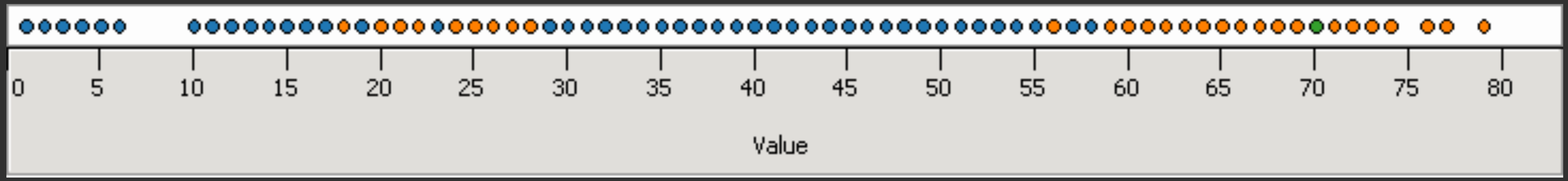
The relationship among multiple data attributes may not be expressed in a single horizontal dot plot.



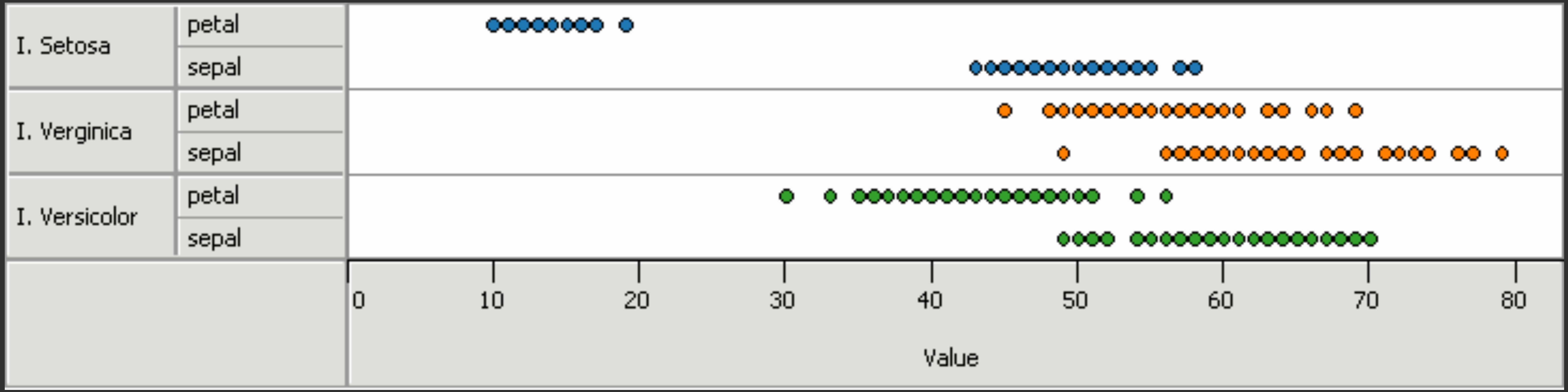
Single horizontal dot plot

Can not express the facts

The relationship among multiple data attributes may not be expressed in a single horizontal dot plot.



Single horizontal dot plot



Categories in different positions

Expresses facts not in the data

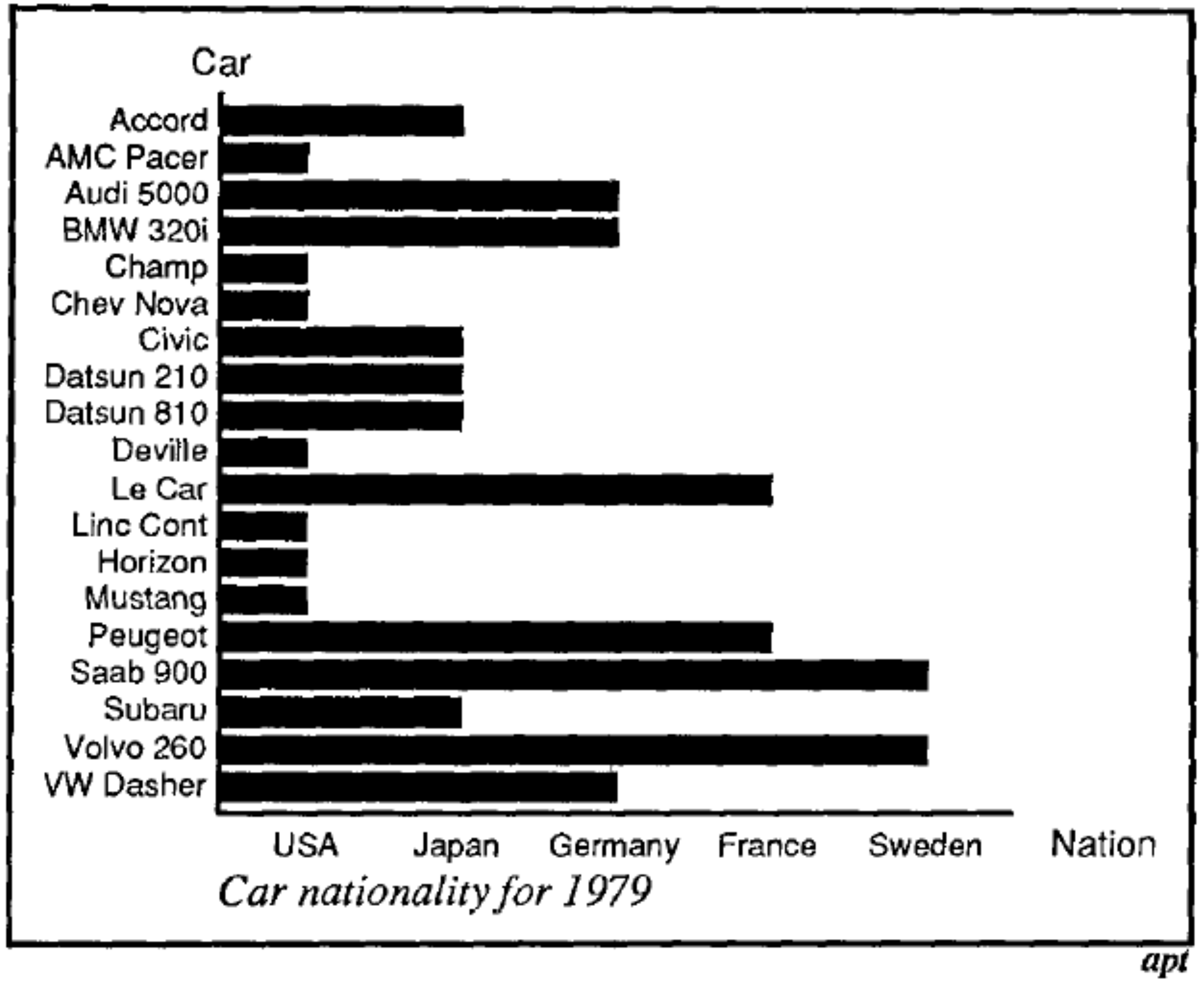


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

A length is interpreted
as a quantitative value.

Design Criteria

Expressiveness

A set of facts is expressible in a visualization if it expresses all the facts and only the facts in the data.

Effectiveness

A visualization is more effective than another one if the information conveyed is more readily perceived.

Design Criteria Translated

Tell the truth and nothing but the truth
(don’t lie, and don’t lie by omission)

Use encodings that people decode better
(where better = faster and/or more accurate)

Mackinlay’s Design Algorithm

APT - “A Presentation Tool”, 1986

User formally specifies data model and type

Input: ordered list of **data variables** to show

APT searches over design space

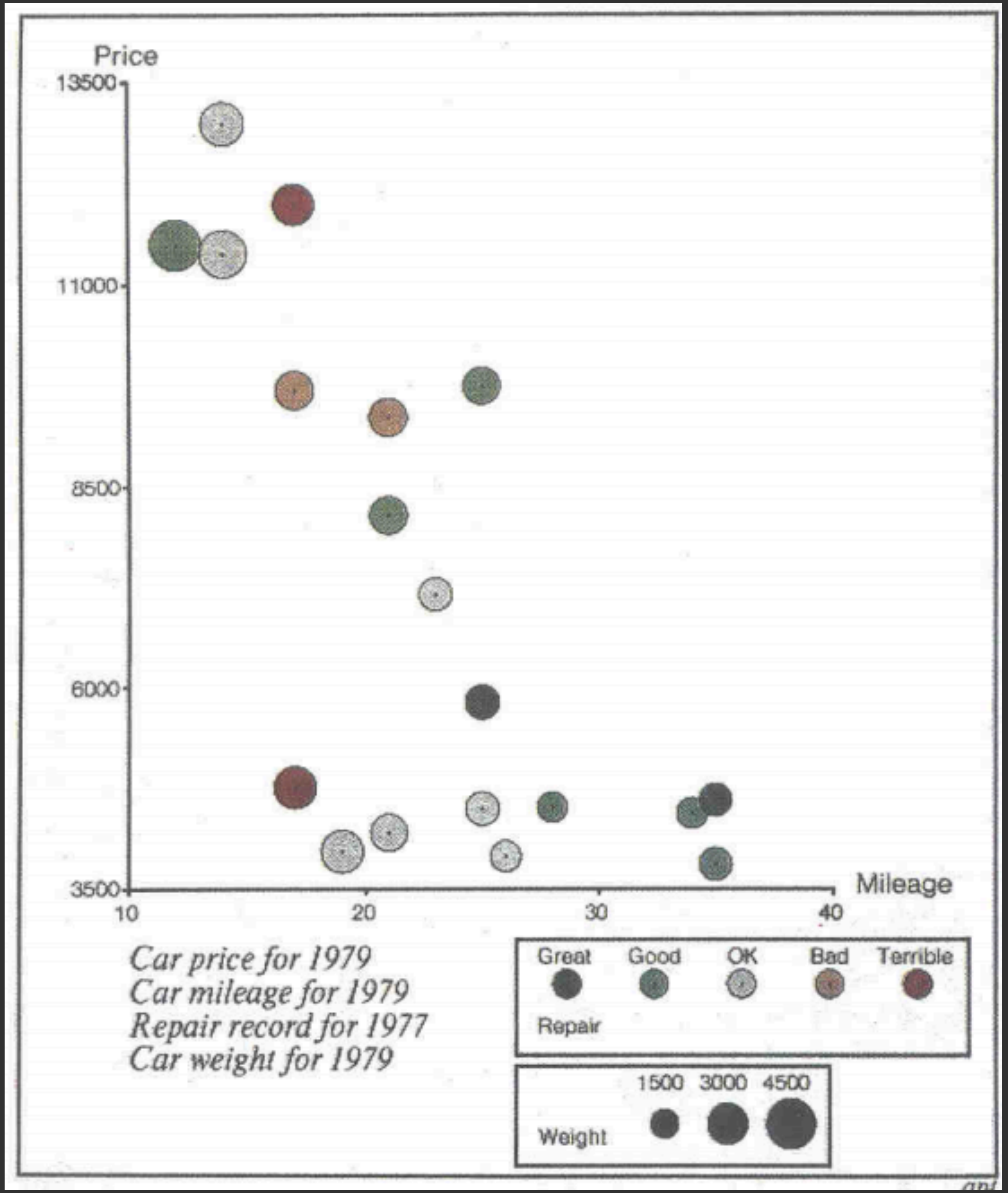
- Test expressiveness of each visual encoding
- Generate encodings that pass test
- Rank by perceptual effectiveness criteria

Output the “most effective” visualization

APT

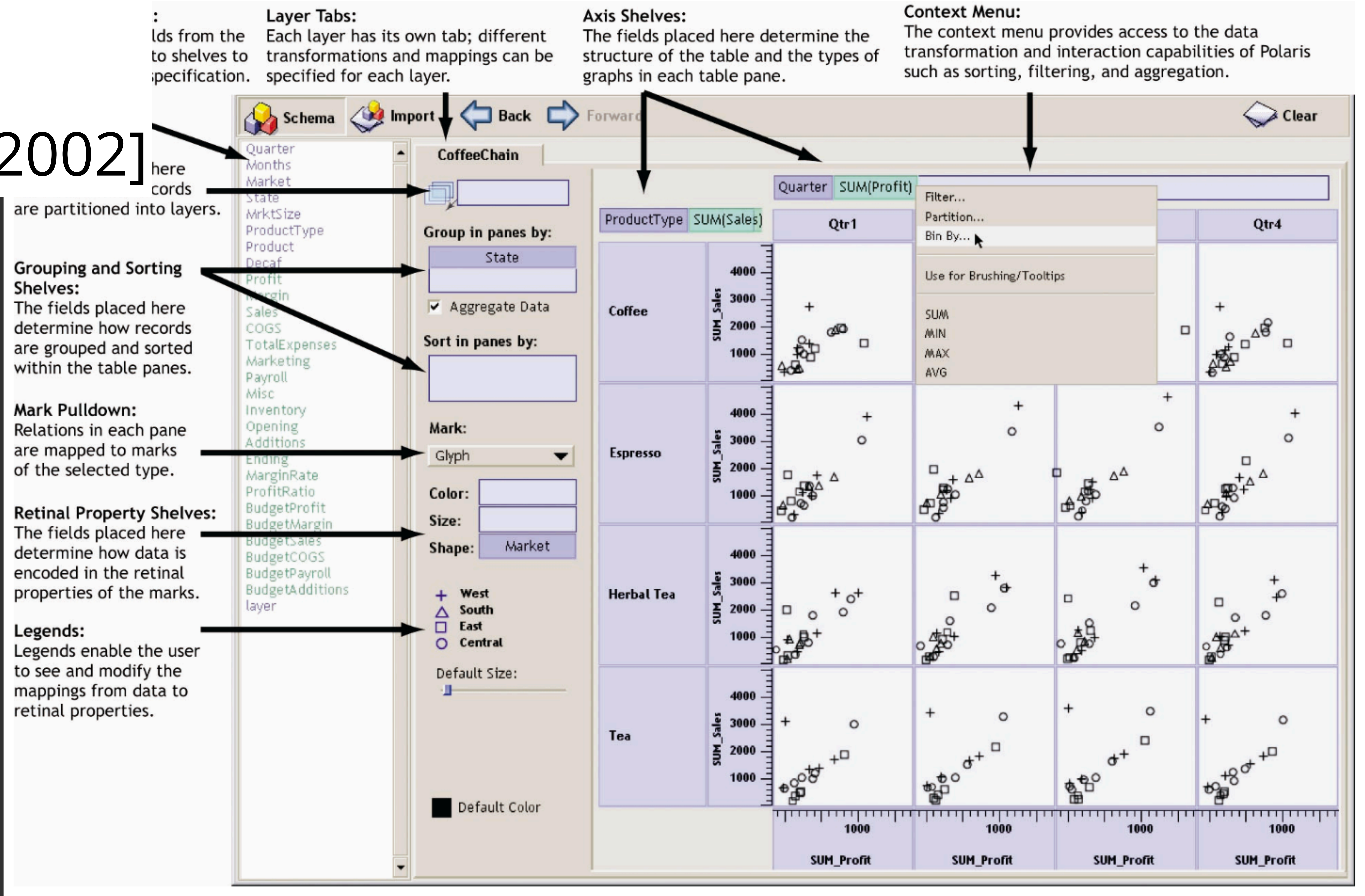
Automatically generate a chart for input variables:

1. Price
2. Mileage
3. Repair
4. Weight



Polaris

[Stolte et al 2002]



:
...lds from the
...to shelves to
...pecification.

Layer Tabs:
Each layer has its own tab; different transformations and mappings can be specified for each layer.

Axis Shelves:
The fields placed here determine the structure of the table and the types of graphs in each table pane.

Context Menu:
The context menu provides access to the data transformation and interaction capabilities of Polaris such as sorting, filtering, and aggregation.

...here
...cords
are partitioned into layers.

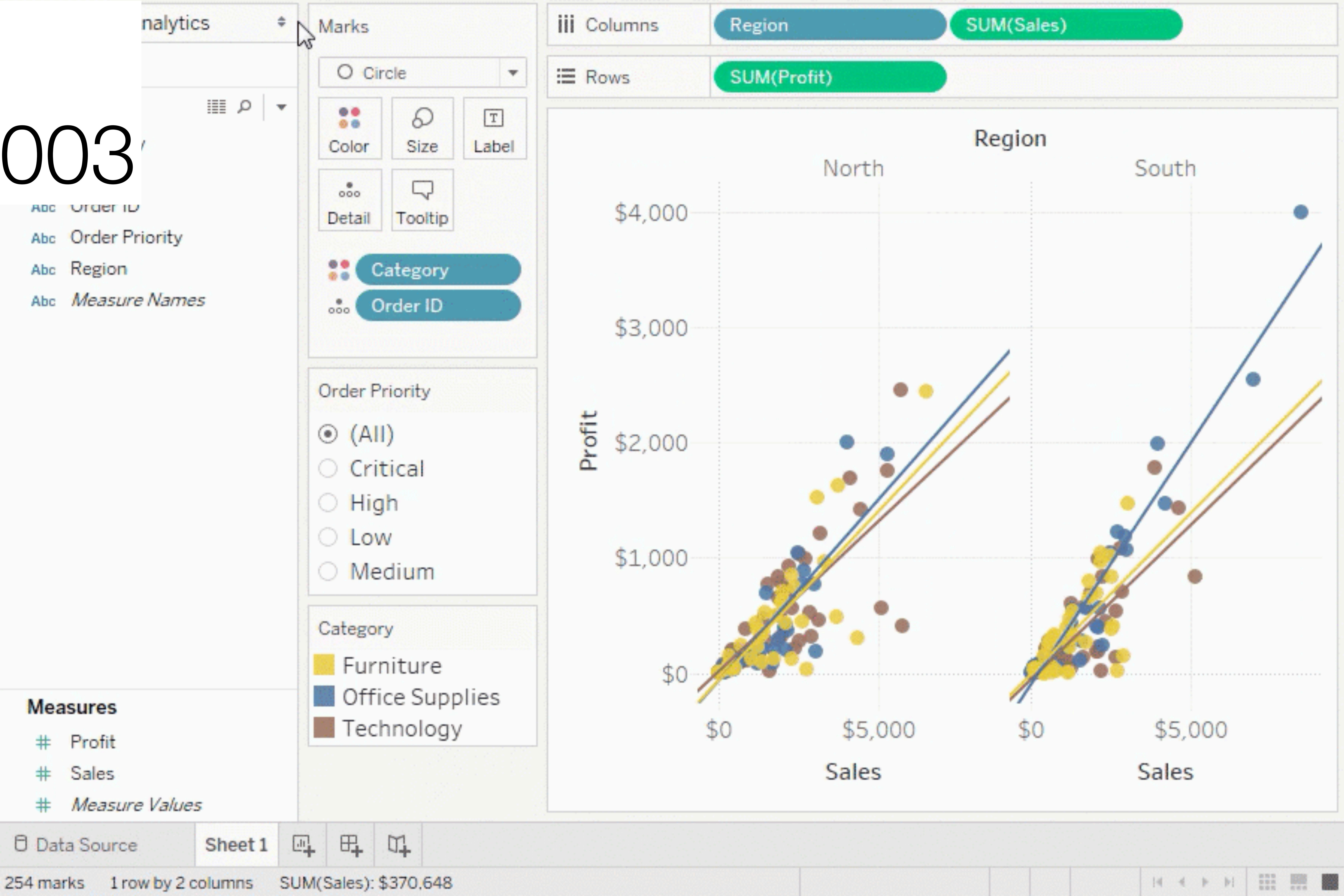
Grouping and Sorting Shelves:
The fields placed here determine how records are grouped and sorted within the table panes.

Mark Pulldown:
Relations in each pane are mapped to marks of the selected type.

Retinal Property Shelves:
The fields placed here determine how data is encoded in the retinal properties of the marks.

Legends:
Legends enable the user to see and modify the mappings from data to retinal properties.

Tableau
founded 2003



Take away: Visual Encoding Design

Use expressive and effective encodings

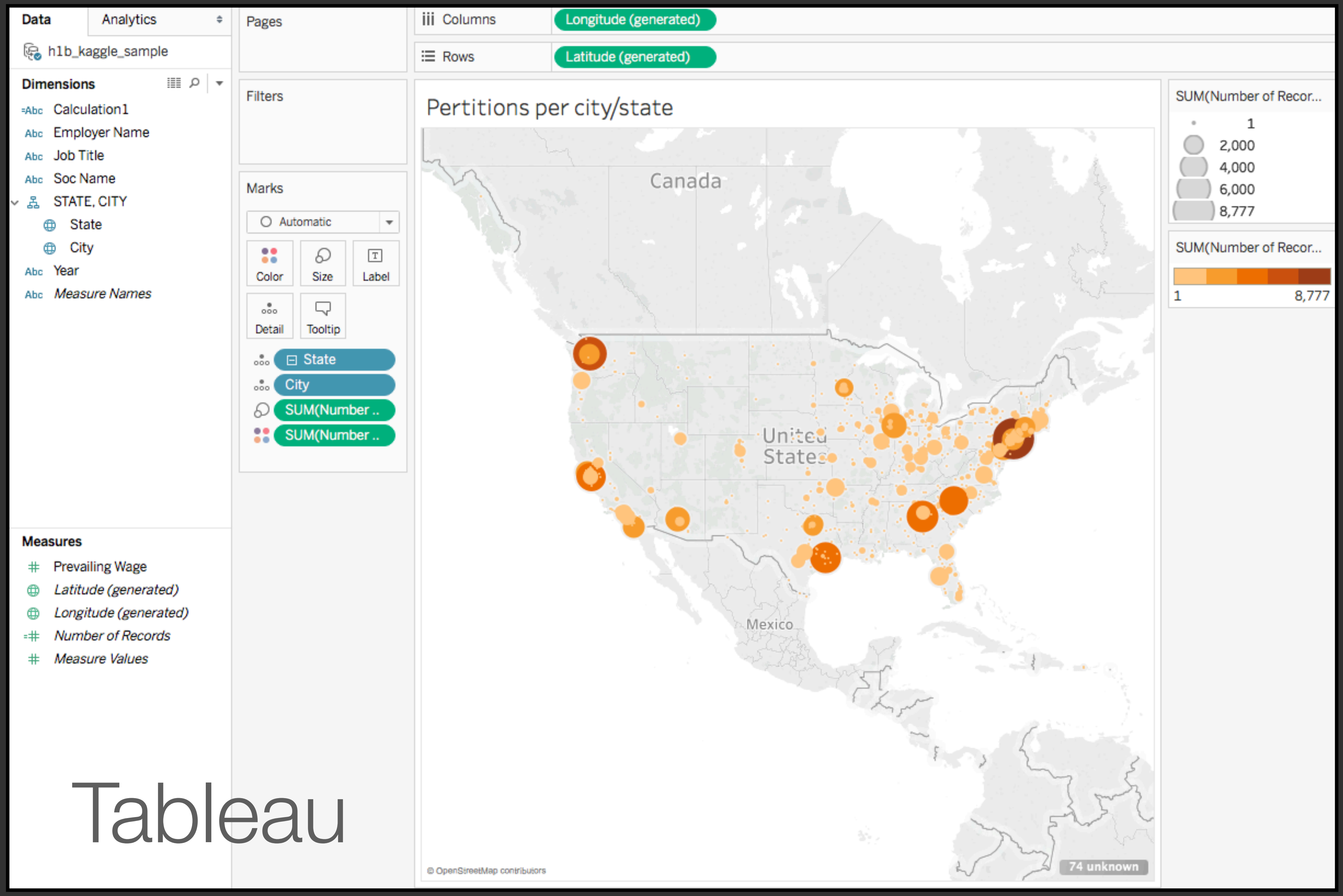
Avoid over-encoding

- Reduce the problem space
- Use space and small multiples intelligently
- Use interaction to generate relevant views

Rarely does a single visualization answer all questions. Instead, the ability to generate appropriate visualizations quickly is critical!

Next

Exploratory Data Analysis



Tableau

H-1B petitions filed in each state

10 min break

Download Tableau & H-1B petition data